Evaluating the Usefulness of Synthetic Data in Healthcare: Applications in Predictive Modeling and Privacy Protection

by

Mohammad Ahmed Basri

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2024

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The advent of data-driven approaches in healthcare has opened new horizons for patient care, disease management, and medical research. However, one of the significant challenges is the availability of large-scale, high-quality datasets. Accessing health data that contains sensitive information requires lengthy approval processes and stringent restrictions. Synthetic data effectively addresses this dilemma by replicating the statistical properties of real datasets, offering a viable solution. Due to privacy concerns and regulatory restrictions associated with health data, there is a growing need for highly realistic synthetic health data, particularly in health data science initiatives. While significant advancements have been achieved in establishing recognized evaluation methods for synthetic data models, there remains a notable gap in understanding the optimal approaches to enhance the quality and usefulness of synthetic data. This thesis aims to bridge this gap by conducting a systematic evaluation of objective functions for hyperparameter tuning of synthetic data generation and studying the efficacy of synthetic data in predictive models. We evaluate synthetic data using three criteria: Fidelity, assessing how well it mirrors real-world data statistically; Utility, measuring its effectiveness for machine learning applications; and Privacy, evaluating the risk of re-identification. We examine the usefulness of synthetic data for the hyperparameter optimization process of predictive models, particularly in scenarios where access to real data is constrained. We found a notable correlation between model performance accuracy using real data and synthetic data, suggesting that parameters optimized with synthetic data are applicable to real data for optimal results. Our study confirms the feasibility of using synthetic data on external computing resources to optimize models, effectively addressing healthcare's computing constraints.

Acknowledgements

I extend my deepest gratitude to my supervisors, Professor Helen Chen and Professor Alexander Wong, for their invaluable support and guidance throughout my journey. Their passion for research is truly inspiring, and their mentorship has significantly contributed to my development as a scholar.

I would like to express my profound appreciation to the Public Health Agency of Canada (PHAC). The collaboration with PHAC has been instrumental in shaping the research that forms the backbone of this thesis. The insights, resources, and innovative perspectives provided by PHAC have enriched my work in ways I had not envisioned at the outset of my research.

I am immensely thankful to the entire team at the Digital Intelligence for Public Health (DI4PH) and the Vision and Image Processing (VIP) Lab. The opportunity to collaborate, learn, and engage in intellectual discussions has been a highlight of my academic experience. Special thanks to labmates Bing Hu, Abu Yousuf and Amy Tai for their support and contributions. I am particularly grateful to Professor Zahid Butt for their insightful guidance on my thesis and other research endeavors.

My appreciation also extends to Professor Sirisha Rambhatla and Professor Abel Torres Espin for their willingness to review my thesis. Your thorough examination and constructive feedback have been instrumental in refining my work.

Lastly, but most importantly, I wish to thank my family and friends. Your unwavering support and belief in my capabilities have been my pillar of strength throughout this challenging and rewarding journey.

Dedication

This is dedicated to my parents, whose unwavering support and encouragement have been the bedrock of my journey.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the growing era of artificial intelligence (AI) and machine learning (ML), the potential for transforming healthcare through advanced applications is immense. These technologies promise to revolutionize disease diagnosis, treatment optimization, patient care, and health system management, underscoring a paradigm shift towards more personalized and efficient healthcare systems. However, the fuel that powers these advanced AI and ML algorithms—data—is often stuck in complexities surrounding accessibility and privacy concerns.

Despite the vast amounts of health-related data generated through Electronic Health Records (EHRs), patient self-reported information, and healthcare service transactions, harnessing this wealth of data for research and development poses significant challenges [1]. The difficulties in leveraging such data do not arise from the lack of it but are primarily due to the rigorous protective measures established to safeguard sensitive and personal health information. These datasets, abundant in valuable insights for advancing healthcare, are subject to comprehensive approval processes and access restrictions [2]. This is to ensure that even de-identified data is utilized in a manner that is secure and ethical, and respects patient confidentiality. The safeguards in place, while necessary for protecting patient privacy, often slow the pace at which health data can be made available for research purposes, thus affecting the speed and scope of advancements in healthcare technology. Regulations like the Health Insurance Portability and Accountability Act (HIPAA) and Personal Information Protection and Electronic Documents Act (PIPEDA) impose strict guidelines to ensure patient data privacy and security, further complicating the research and development process [3].

The emphasis on data sharing and accessibility has never been more critical, especially

with the growth of digital healthcare [4]. Open data initiatives have increasingly recognized health data as a cornerstone for fostering innovation and enhancing the efficacy of AI and ML applications in healthcare. These initiatives aim to navigate the delicate balance between making data broadly available to researchers and developers and maintaining the inviolability of patient privacy [5]. The importance of this equilibrium cannot be overstated, as it directly impacts the pace of technological advancements and the realization of AI's full potential in healthcare.

Moreover, the protection of data privacy and the implications of confidential patient information breaches underscore the ethical dimensions of data accessibility. According to a 2020 study on 'Healthcare data breaches', the rate of data breaches has increased even more rapidly in the last five years [6]. The repercussions of such breaches extend beyond individual privacy violations, potentially eroding public trust in healthcare systems. Thus, as we advance into a future where AI and ML play pivotal roles in healthcare, addressing these challenges becomes paramount.

Synthetic health data emerges as a promising solution to the delicate balance between protecting patient privacy and maintaining the utility of data for advancing ML applications in healthcare. By generating data that mirrors real patient information in structure and statistical properties, yet does not correspond to any actual individuals, synthetic data holds the potential to revolutionize patient care. This innovative approach not only safeguards against privacy breaches (by enabling secure data sharing) but also offers a rich dataset for researchers and developers to train and refine AI models, thereby accelerating the pace of healthcare innovation without compromising on confidentiality.

This work aims to explore the utilization of synthetic health data as a means to navigate the intricacies between data accessibility and privacy preservation. Specifically, the focus is on assessing the effectiveness of synthetic data for predictive modelling in healthcare applications.

1.1 Problem Definition

While significant progress has been made in the development of recognized evaluation methodologies for synthetic data generation models, a deficiency exists in pinpointing the most effective strategies and objective functions for hyperparameter optimization, which is critical for enhancing the quality of synthetic health data. Additionally, much of the existing literature and frameworks are designed to optimize synthetic data performance within narrow, predefined use cases rather than for general applicability across diverse healthcare

analytics applications. To tackle this we develop a methodology for Hyperparameter Tuning (HPT) of Synthetic Data Generation (SDG) for use-case agnostic synthetic data by optimizing different quality metrics along with multi-objective optimization.

Historically, the utilization of synthetic health data has predominantly been focused on two major use cases: 1) privacy-enhancing technique and 2) data augmentation, serving as a mechanism to augment the robustness of predictive analytics models. However, the broader applications of synthetic health data, particularly in its utility of building highly realistic machine models and model optimization, remain underexplored. Prior research has highlighted the importance of hyperparameter tuning in maximizing the efficacy of predictive analytics models [7]. Entities with access to voluminous healthcare datasets frequently encounter computational resource constraints, impeding their capacity for extensive model training and the computationally very expensive hyperparameter optimization necessary for achieving superior model performance. The second phase of this study delves into conducting hyperparameter optimization on synthetic data for subsequent application to real-world datasets through a series of experiments using diverse datasets, a broad spectrum of models, and various scenarios. This methodical approach allows us to assess the viability and impact of leveraging optimized synthetic data in enhancing models' performance and predictive accuracy when applied to genuine healthcare data. The use of synthetic data in explaining the models trained on real data has also been explored by comparing the explainability of real and synthetic data-trained models.

1.2 Contributions and Outline

The objective of this thesis is to advance open data science efforts in the healthcare sector through a comprehensive assessment of hyperparameters to enhance the quality of synthetic data, as well as evaluate the effectiveness of utilizing synthetic data during the hyperparameter optimization phase of ML algorithms.

Given the gap in research and the importance of improving synthetic data generation to enhance predictive model performance, this thesis focuses on the following key questions. Firstly, how can objective functions for hyperparameter tuning be systematically evaluated to improve the generation of synthetic data? Secondly, how can we assess the quality of synthetic data using fidelity, utility, and privacy metrics, and what is the impact of these measurements on the data's overall effectiveness? Lastly, considering the constraints on access to real data and computational resources, how can synthetic data be optimized to enhance the hyperparameter optimization process of predictive models?

The rest of the thesis is organized as follows: Chapter 2 provides background information about AI and ML applications in healthcare, synthetic data generation methods, and their evaluation metrics. Chapter 3 details datasets used, their preprocessing and the model used for generating synthetic health tabular data. It also includes the methodology for hyperparameter tuning of synthetic data generation and use of synthetic data for hyperparameter tuning of predictive models. Results and their implications are discussed in Chapter 4. Chapter 5 will summarize the main findings and impact of this research. In the same section, we also present future works. Some of the code supporting this research is available at the GitHub repository: Usefulness-of-Synthetic-Health-Data.

Chapter 2

Literature Review

The advent of data-driven approaches in healthcare has opened new horizons for patient care, disease prediction, and medical research. However, one of the significant challenges in this domain is the availability of large-scale, high-quality datasets. Due to privacy concerns and regulatory restrictions associated with patient data, there is a growing need for synthetic data that can replicate the statistical properties of real healthcare datasets without compromising patient privacy. This thesis explores the various methodologies employed in generating synthetic tabular data for healthcare applications, a field that stands at the crossroads of data privacy, machine learning, and healthcare innovation.

The healthcare industry generates vast amounts of data, ranging from patient records and clinical trial data to disease registries and insurance claims. This data is inherently tabular, with rows representing individual patients and columns corresponding to various attributes, such as demographic details, diagnoses, treatment information, and outcomes. The potential of this data is immense, particularly for training machine learning models that can predict patient outcomes, personalize treatments, and improve healthcare delivery. However, the sensitive nature of this data imposes stringent constraints on its accessibility and use, thus necessitating the creation of synthetic datasets that are realistic enough to be useful but do not contain any real patient information.

This thesis delves into the core methodologies for generating synthetic tabular data, with a particular focus on their application in the healthcare sector. It critically analyzes the effectiveness of various approaches, evaluates their performance in terms of data fidelity and privacy preservation, and discusses their practical implications in healthcare analytics and research. Through a comprehensive review and experimental investigations, this work aims to contribute to the evolving field of synthetic data generation, offering insights and

guidance for researchers and practitioners in healthcare data science.

2.1 Applications of AI and ML in Health and the Importance of Data

Recently there has been a tremendous amount of growth in machine learning. A major field of application of AI and ML is for health and medical data. With the tremendous growth in this field, researchers have taken leverage of ML and DL models for the abundant and complex medical data [8]. In the rapidly evolving field of healthcare, Artificial Intelligence (AI) and Machine Learning (ML) technologies are revolutionizing how we collect, analyze, and interpret a wide variety of data types to improve patient outcomes, streamline operations, and unlock new insights. From the detailed imagery capturing the intricacies of human anatomy to the rich genetic data offering a blueprint of life, each data type presents unique challenges and opportunities for innovation. Electronic Health Records (EHR) provide a digital history of patient interactions with healthcare systems, encapsulating a wealth of information for analysis. Meanwhile, the advent of wearable technology and Internet of Things (IoT) devices has unleashed plethora of real-time health data, offering unprecedented monitoring capabilities outside traditional healthcare settings. Additionally, the vast, unstructured datasets from social media and online interactions offer a new lens through which to view public health trends and patient wellness. Together, these diverse data streams form the backbone of AI and ML applications in health, each requiring specialized approaches to harness their full potential for advancing healthcare.

Convolutional Neural Network (CNN) has been extensively used with image-based data for cancer detection. [9]. Spanhol et al. effectively utilize the AlexNet CNN architecture to classify breast cancer images from the BreaKHis dataset, achieving superior accuracy over traditional models without the need for hand-crafted features [10]. This approach demonstrates the potential of repurposing existing CNN models for complex medical image analysis. Paul et al. explores lung tumor feature extraction using transfer learning from a pre-trained CNN, significantly improving lung cancer survival time prediction accuracy beyond traditional methods [11]. This study underscores the benefit of combining deep learning with quantitative analysis in cancer prognosis. Zhao et al. presents innovative CNN architectures for brain tumor segmentation, adeptly handling multimodal MR images and surpassing existing segmentation methods [12]. By converting the 3D challenge into a 2D analysis, this work showcases the adaptability and efficiency of CNNs in medical imaging, paving the way for future advancements in multimodal image processing and tumor classification. Together, these studies highlight the transformative impact of CNNs

and deep learning techniques in advancing medical imaging and diagnostics across various domains.

The integration of machine learning (ML) with tabular health data has opened new avenues in healthcare, transforming how patient information is analyzed and utilized for clinical decision-making. By leveraging structured datasets that encapsulate diverse patient attributes, such as demographics, laboratory results, and medical histories, ML algorithms can uncover intricate patterns and relationships. This capability not only enhances the accuracy of diagnoses and prognoses but also tailors preventive and therapeutic interventions to individual patient needs [13]. Consequently, the synergy between ML and tabular health data is instrumental in advancing personalized medicine, optimizing healthcare delivery, and ultimately improving patient outcomes. Chang et al investigated the prediction of outcomes in hypertensive patients using a novel method that combines classifiers (Support Vector Machine, Decision Tree, Random Forest, XGBoost) with recursive feature elimination with cross-validation, focusing on essential physical examination indicators [14]. Their study revealed that indicators like limb and ambulatory blood pressure play significant roles in hypertension outcomes. The research highlighted XGBoost's superior prediction accuracy and its potential for telemedicine applications, enabling efficient, targeted interventions for patients at higher risk. In their work, Rahimian et al presented an improvement in predicting the risk of emergency hospital admissions, showing machine learning models' superiority over traditional statistical models by enhancing model discrimination and calibration [15]. The study emphasized the value of adding variables and timing information to the models, which maintained high performance over longer prediction windows. This advancement in using EHR data for predictive modeling is based on a dataset of 4.6 million patients, showcasing a significant leap forward in healthcare analytics. Liao et al aimed to improve the management of COPD patients by developing prediction models for acute respiratory failure, ventilator dependence, and mortality post-hospitalization, using machine learning algorithms [16]. The models, validated on a dataset of 5061 COPD patients from three hospitals, demonstrated excellent predictive quality with algorithms like XGBoost, random forest, and LightGBM. The integration of the best models into a web service application for hospital use underscores the potential of these algorithms in supporting physicians' decision-making processes. Miotto et al introduced "Deep Patient," a deep learning approach for deriving predictive patient descriptors from EHRs [17]. Their method outperformed traditional feature learning models and the use of raw EHR data by capturing complex data patterns through deep sequence of nonlinear transformations. This approach not only enhances disease prediction but also offers a scalable solution to the growing volume of hospital data, marking a significant contribution to the field of clinical predictive modeling. Jaotombo et al explored the use of Machine

Learning models to predict Prolonged Hospital Length of Stay (PLOS), identifying the GB classifier as the most effective based on a dataset of 73,182 hospitalizations [18]. The study pinpointed the destination of the patient post-hospitalization as a critical predictor of PLOS, providing valuable insights into the risk profile of elderly patients with specific conditions. This research contributes to the ongoing efforts to improve healthcare quality and efficiency by identifying key predictors of PLOS.

2.1.1 Privacy in Health Data Management

The imperative for stringent privacy measures in healthcare data management highlights a critical aspect of modern medical practice, balancing the need for data accessibility with the protection of individual privacy rights. In the context of growing public interest in healthcare quality, cost, and accessibility, the Institute of Medicine (IOM) underscores the urgent need for enhanced privacy protection in health data management [19]. With the digital era enabling vast data collection and management opportunities, the IOM advocates for a federal statute to standardize privacy protections, superseding varied state laws. This proposed legislation would not only ensure data confidentiality across state lines but also establish a Code of Fair Health Information Practices, allowing individuals to review and dispute their health information. It emphasizes the importance of perceiving data sensitivity based on potential harm, advocating for universal stringent data protection.

Chong et al. present a thorough investigation into the burgeoning realm of Electronic Health Records (EHRs), emphasizing their critical role in enhancing healthcare service delivery while highlighting the concomitant privacy challenges associated with the publication of sensitive patient data [20]. It delves into state-of-the-art privacy-enhancing methodologies, with a particular focus on data anonymization and differential privacy techniques, designed to ensure the secure sharing of healthcare data. These methods are critically examined for their strengths and limitations, offering a roadmap for navigating the complexities of healthcare data privacy. Amidst the exploration of data collection and publication processes within the healthcare sector, Chong et al. categorize data attributes into explicit identifiers, quasi-identifiers, and sensitive, and non-sensitive attributes [20].

- **Identifiers (ID):** Attributes that can uniquely pinpoint the identity of a patient. These include, but are not limited to, an individual's name, Social Security number, national identification numbers, mobile phone numbers, and driver's license numbers.
- **Quasi-Identifiers (QID):** Attributes that, on their own, do not reveal the identity of a patient but could do so when combined with other information. Examples are date of birth, gender, residential address, postal code, and personal interests.

- **Sensitive Attributes (SA):** Personal data that individuals prefer to keep confidential. This category encompasses medical diagnosis codes, genetic data, financial earnings, specific health conditions, insurance details, and social relationships.
- **Non-Sensitive Attributes (NSA):** Information that, when disclosed, does not compromise the privacy of the individual. This includes information such as cookie identifiers, anonymized email addresses, and mobile advertising identifiers derived from Electronic Medical Records (EMR).

This classification underpins the discussion on the necessity of Privacy-Preserving Data Publishing to prevent unauthorized access to personal patient information.

Privacy disclosure emerges as a critical concern in the realm of healthcare data publication, defined as the unauthorized release of personal information that individuals wish to keep confidential [20]. This phenomenon manifests in three distinct forms: identity disclosure, attribute disclosure, and membership disclosure. Identity disclosure, or reidentification, represents a significant risk, occurring when an adversary can match a published record to its corresponding individual with notable accuracy, effectively unveiling the person's identity. Attribute disclosure involves the unauthorized linkage of a person to specific confidential attributes, such as medical conditions, disclosed within the data. Membership disclosure, on the other hand, pertains to the ability of an adversary to ascertain an individual's presence within a particular dataset, such as identifying someone within a database of individuals tested positive for a disease. Each of these disclosure types underscores the paramount importance of implementing robust privacy-preserving mechanisms in the dissemination of healthcare data to mitigate the risks of personal information exposure.

The Health Insurance Portability and Accountability Act of 1996 (HIPAA) introduced the Privacy Rule, targeting health providers, plans, and clearinghouses by standardizing health information privacy practices [3, 21]. While the rule limits how covered entities can use health information, it allows health plans to require authorizations for enrollment and does not restrict noncovered entities like employers and life insurers from demanding authorizations for employment or insurance purposes. The Personal Information Protection and Electronic Documents Act (PIPEDA) is a Canadian federal law that governs how private sector organizations collect, use, and disclose personal information in the course of commercial business [22]. Enacted in April 2000, PIPEDA aims to balance an individual's right to privacy with the need for organizations to use personal information for legitimate business purposes. This delineates a framework for the privacy and authorization of health information, balancing patient privacy with operational necessities. In Canada, the implementation of PIPEDA for health data is under the competence of the provinces, with

examples like the Personal Health Information Protection Act (PHIPA) in Ontario. Such a regulatory framework can generate challenges for healthcare research, particularly in scenarios involving cross-province data linkage. In this context, the utility of synthetic data becomes even more significant.

2.2 Generation and Application of Synthetic Data

De-identification or anonymization is the process of removing or altering personally identifiable information from data sets so that the people whom the data describes remain anonymous. However, this approach alone is insufficient to guarantee the privacy of individuals in real-world health data. De-identified datasets can often be susceptible to re-identification through methods such as linkage attacks, which leverage external data sources to match anonymized records to specific individuals. The distinct patterns and attributes within health data can inadvertently serve as identifiers, especially when dealing with rare conditions or demographic specifics. The growing availability of varied datasets and advancements in data analytics exacerbate the vulnerability of de-identified data to privacy breaches.

Despite a growing interest in synthetic data, there is a lack of a unified definition of what represents synthetic data. Earlier definitions defined synthetic data as the generation of novel data points that reflect the statistical features of the source data. The Royal Society alongside The Alan Turing Institute has recently suggested a provisional definition, characterizing synthetic data as the outcome of data generated by specifically designed mathematical models or algorithms, intended to address certain data science challenges [23].

Synthetic data can accelerate research by enabling quicker access to sensitive data [24]. Synthetic health data not only enhances data accessibility but also significantly streamlines the research timeline, as depicted in Figure 2.1. It reveals that when synthetic data is employed, the data analysis timeline, including data governance (encompassing ethical and legal considerations), and data preprocessing is substantially shortened. This compression of the timeline enables researchers to swiftly transition to the data preprocessing and analysis stage. As a result, researchers can dedicate more time to the analytical work itself, fine-tuning their methods and ultimately expediting the journey to the Results. When delays in accessing real data occur—a common occurrence given the stringent privacy protections required—researchers equipped with synthetic data can develop comprehensive analysis plans, write analysis code, and troubleshoot potential issues in advance. However,

it is important to acknowledge that the initial setup for creating synthetic data involves additional time and resources for model development and data generation. Furthermore, the need for periodic updates to the models to ensure the synthetic data remains representative of current datasets adds another layer of complexity.

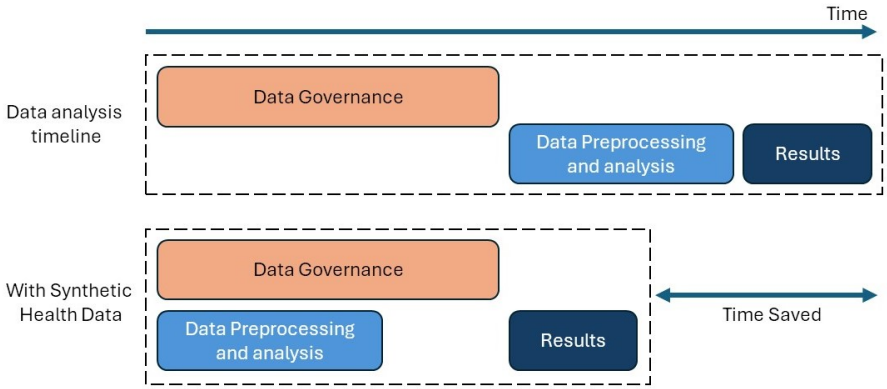


Figure 2.1: Impact of using Synthetic Data to Accelerate Data Analysis

Synthetic data can be utilized in various healthcare tasks to enhance machine learning algorithms, including those used in image classification workflows. It can also serve as a foundational basis for pre-training these models. These pre-trained models can subsequently be refined for distinct patient demographics. Additionally, synthetic data can bolster public health models, aiding in the prediction of infectious disease outbreaks [25, 26, 27].

In a 2020 study, Julia et al. showcased the use of synthetic data in training a Natural Language Processing (NLP) model with datasets derived from patient discharge notes [28]. This approach proved effective in identifying and predicting mental health conditions and phenotypes. By using Electronic Health Records (EHRs) for NLP data, the model could pinpoint key disease characteristics and progressions. Specifically for mental health, leveraging synthetic unstructured text helps train models on complex conditions while safeguarding patient privacy, as mental health data is particularly sensitive.

Synthetic data is instrumental in advancing digital twins within healthcare, offering a way to create customized patient models that enhance treatment optimization and patient outcomes [29, 30]. Its application extends to improving hospital efficiency and operations, allowing for the simulation of various scenarios such as changes in patient intake, staff competency, and equipment access. This enables effective resource allocation and staffing adjustments, leading to cost reductions and better patient care.

Synthetic health data serves as a valuable resource for data augmentation, significantly enhancing the robustness and generalizability of machine learning models in healthcare. By generating diverse and realistic datasets, synthetic data helps to overcome limitations posed by small or imbalanced real-world datasets, facilitating more comprehensive training environments [31, 32]. This approach not only improves model performance but also aids in achieving better diagnostic accuracy and predictive analytics in varied clinical scenarios.

The burgeoning field of synthetic data generation has witnessed a variety of models and approaches, particularly geared towards tabular data, which is predominant in healthcare datasets. An ideal Synthetic data generation models learn complex multidimensional patterns from real data by training on the underlying distributions and features, allowing them to generate new samples based on the learned parameters of the model.

Early methods for privacy protection relied on anonymization, replacing values and removing sensitive variables. These methods did not use data modelling for data generation [33]. Simple random sampling and bootstrapping techniques were foundational but often struggled to capture complex correlations between variables, a challenge in the multidimensional space of healthcare data.

More recently, advanced machine learning techniques, including generative adversarial networks (GANs) and variational autoencoders (VAEs), have shown promise in generating high-fidelity synthetic data [34, 35, 36]. These methods leverage the power of deep learning to model the intricate relationships within healthcare datasets, offering a balance between data utility and privacy.

Traditional machine learning models offer a pragmatic approach to generating synthetic tabular data, blending the benefits of simplicity and effectiveness. These models, including decision trees, Bayesian networks, and k-nearest neighbours, are adept at capturing the intricate relationships within datasets, ensuring the synthetic data maintains statistical fidelity to the original. Their versatility and capacity for integrating privacy-preserving techniques make them invaluable tools in the creation of synthetic datasets across various domains.

Within the context of synthetic data generation techniques, the Classification and Regression Trees (CART) method is employed to create anonymized health datasets by constructing decision trees that model the relationships between variables. This method can manage both continuous and categorical data, making it suitable for the multifaceted nature of health data. The primary function of CART in synthetic data generation is to facilitate the derivation of datasets that can be used for research while ensuring individual data points remain non-identifiable [34].

Bayesian networks model the probabilistic relationships among variables and have been

used to generate synthetic datasets that preserve the statistical properties of the original data [35]. The conditional probabilities inherent in Bayesian networks can adeptly handle the intricacies of healthcare data, which often includes a mix of discrete and continuous variables.

Among the deep learning methods of synthetic data, the most commonly used is a Generative Adversarial Network (GAN) model. The adversarial process within GANs, involving a generator and a discriminator, has proven adept at generating complex, high-dimensional data [37]. Adaptations of GANs for healthcare data have addressed the challenge of generating synthetic patient records that are both realistic and diverse.

One of the initial works on using GAN for Electronic Health Records (EHR) was by Choi et al. in their work medGAN [36]. medGAN represents an innovative approach to addressing the issue of mode collapse in the generation of medical binary and categorical data. By integrating an Autoencoder (AE) with a Generative Adversarial Network (GAN), this model effectively manages binary variables. Additionally, it incorporates a minibatch averaging technique and batch normalization to enhance its performance and stability. Park et al. developed the table-GAN method which builds upon DCGAN (Deep Convolutional GAN). Table-GAN utilizes a Discriminator (D) structured as a Convolutional Neural Network (CNN) with various layers, applying a series of 3x3 learning filters throughout. The final layer employs a sigmoid activation function while preceding layers benefit from batch normalization and the use of LeakyReLU activation functions. Conversely, the Generator (G) is constructed from de-convolutional layers within a neural network framework, utilizing a loss function termed 'information loss' [38]. The novel architecture of Sequentially Coupled GAN (SC-GAN) incorporates dual generators, intricately linking the generation of patient states with that of medication dosages to reflect their real-world interdependence. This model is intricately designed with a two-layer bidirectional LSTM for the Discriminator (D), and two separate two-layer LSTMs for generating patient status and medication dosage, respectively. The medication dosage generator is particularly innovative, synthesizing data based on the sequential patient status and injected noise, while the patient-status generator integrates previous states, medication dosages, and noise. Empirical evaluation on actual medical tasks demonstrates SC-GAN's superiority in generating data that enhances model performance over other generative approaches, highlighting its potential in circumventing privacy concerns and data scarcity in healthcare analytics [32].

The Wasserstein distance loss function solves the mode collapse problem in GANs. The Wasserstein GAN (WGAN) model [39] has been used in various works, along with some modifications for tabular health data. Chin-Cheong et al explores the use of Generative Adversarial Networks (GANs) to generate synthetic, heterogeneous Electronic Health Records (EHRs), incorporating differential privacy (DP) to enhance data privacy and shareability

[40]. The authors report that while synthetic data without DP closely mirrors the original data within a 6.4% margin, DP synthetic data exhibits a 20% performance drop but remains viable for machine learning purposes. Using a Wasserstein GAN (WGAN) approach, the research demonstrates the model's ability to accurately replicate EHR statistical properties, highlighting its potential for creating privacy-compliant, synthetic healthcare datasets with minimal divergence from real data characteristics. Another implementation of WGAN is HealthGAN, a Generative Adversarial Network designed for generating synthetic health data that meets privacy and quality needs for education and research [41]. Through a unique set of evaluation metrics, HealthGAN is shown to excel in producing data that closely resembles real datasets, while ensuring privacy and maintaining utility and compactness. Its training within a secure environment and subsequent availability for external generation offers a practical solution to the challenges of data de-identification.Health-GAN claims to solve the compatibility and divergence problems of medGAN.

The Synthetic Data Vault (SDV) emerges as a pioneering framework in the realm of synthetic data generation, designed to address the pressing need for privacy-preserving synthetic datasets across various sectors [42]. Within the SDV, the Conditional Tabular Generative Adversarial Network (CTGAN) is featured as a key synthetic data generation model [43], specifically tailored to tackle the complexities of tabular data which is prevalent in fields such as healthcare and finance. CTGAN distinguishes itself by adopting a unique training methodology that conditions on the column-wise distribution of data, thereby capturing the intricate, multi-modal nature of real-world data distributions more effectively. This approach enables the generation of synthetic data that is not only diverse and accurate but also maintains the statistical characteristics of the original datasets. By integrating CTGAN within its suite of generative models, the Synthetic Data Vault significantly enhances the ability to produce high-fidelity synthetic datasets, thereby facilitating research and development activities while strictly adhering to data privacy and ethical standards

Differential privacy is a technique designed to ensure that the privacy of individuals in a dataset is protected, such that the output of any analysis does not allow an observer to reliably determine whether any one individual was included in the original dataset. The application of differential privacy to synthetic data generation added a robust privacy-preserving layer to the process. The Renyi-differentially private-GAN (RDP-GAN) advances the field of generative adversarial networks by focusing on privacy protection without sacrificing sample quality [44]. By introducing random Gaussian noises to the loss function during training, RDP-GAN offers a novel approach to ensure differential privacy. This method not only enhances the privacy of sensitive data, such as medical or financial records, but also maintains model stability and eliminates the need for parameter value clipping. The

effectiveness of RDP-GAN is demonstrated through its ability to generate high-quality samples while achieving superior privacy levels compared to traditional DP-GAN models that rely on gradient noise perturbation.

2.3 Evaluation Metrics for Synthetic Data

In the realm of synthetic data generation within healthcare, assessing the quality of synthetic data requires a multifaceted approach, including utility, privacy, and fidelity metrics. Utility measures, such as propensity scores and classification accuracy, examine the synthetic data's ability to preserve the statistical properties and analytical patterns of the original dataset, enabling meaningful research [45]. Privacy metrics, particularly the risk of re-identification, are crucial for evaluating the synthetic data's effectiveness in safeguarding individual identities, a critical aspect of healthcare data management [33]. Fidelity, meanwhile, assesses the accuracy with which synthetic data mirrors the statistical characteristics of the original dataset, ensuring it reflects the real-world complexities and distributions without introducing biases [46].

Focusing on fidelity within the context of synthetic health data, the evaluation involves two primary measures: univariate and bivariate [31]. The univariate approach emphasizes maintaining the distinct distributions of each attribute as found in the original dataset, through a detailed column-by-column comparison between the original and synthetic data. On the other hand, the bivariate measure expands this analysis by examining the associations between pairs of attributes, thereby adopting a pairwise approach to assess the inter-attribute relationships and ensure a more comprehensive fidelity evaluation.

The most common univariate Fidelity metric used for synthetic data is Hellinger Distance [47, 48]. Hellinger distance (HD) quantifies the similarity between two probability distributions [49]. Quantifiable and standard measures such as HD provide open data policymakers additional context alongside visual comparisons of real and synthetic data probability distributions. Bivariate measures are also important considering the fact that the synthetic data must preserve the interdependency between variables. Differential Pairwise correlation (DPC) is a bivariate fidelity metric that measures the strength of correlation between the variables for the real data and synthetic data [50]. Multivariate measures like Kullback-Leibler (KL) Divergence [51] is employed to measure the similarity between the probability distributions of real and synthetic datasets, providing insights into how well the synthetic data represents the underlying statistical properties of the original data. The Log-cluster [51] method evaluates the clustering structure within the data, comparing the log-likelihood of real and synthetic datasets being generated from the same cluster

distribution, thus assessing the preservation of data patterns and groupings. Propensity scores [52], often used in observational studies for causal inference, can be adapted to measure the likelihood of each record belonging to the synthetic dataset versus the real dataset, effectively evaluating the balance and representativeness of the synthetic data. The Kolmogorov-Smirnov (KS) Type Statistic [53] is a non-parametric test that compares the distributions of individual variables between the real and synthetic datasets, identifying discrepancies in cumulative distribution functions to pinpoint variables that may not be accurately modelled.

Utility is a measure employed to evaluate the effectiveness of synthetic data in predictive modeling and analysis. This concept revolves around assessing whether synthetic data can substitute real data for the purpose of developing predictive models. This involves training predictive models using both real and synthetic data, then comparing their performances. Current research in the field of synthetic data highlights three primary methods for gauging synthetic data utility, as discussed by El Emam et al. [47] and El Emam et al. [45]: workload-aware evaluation, generic utility measurement of data, and subjective utility assessments. Specifically, within workload-aware evaluation, the focus is on gauging the suitability of synthetic data for predictive machine learning models, an approach referred to as Machine Learning Efficacy (MLE). In exploring the MLE of synthetic data, various machine learning models are utilized, including XGBoost, Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Trees, and Random Forest.

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient and flexible. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting that solves many data science problems in a fast and accurate way [54]. The key features of XGBoost include handling sparse data, tree pruning, and an efficient implementation of the gradient boosting algorithm. Logistic Regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. [55]. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). It is used for prediction of the probability of occurrence of an event by fitting data to a logistic curve. K-Nearest Neighbors is a simple, versatile, and easy-to-implement supervised machine learning algorithm used for both classification and regression [56]. It works by finding the k-nearest data points in the training dataset to the point that needs to be predicted and then infers its classification or value from these nearest neighbors. KNN makes predictions based on how its neighbors are categorized. Support Vector Machine is a powerful and versatile supervised machine learning algorithm used for both classification and regression challenges [57]. SVM works by finding the hyperplane that best divides a dataset into classes. It is particularly useful for high-dimensional data and is known

for its accuracy and the ability to handle non-linear data. A Decision Tree [58] is a non-parametric supervised machine learning algorithm which has a flowchart-like tree structure where an internal node represents a feature, the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value. It partitions the tree in recursively manner call recursive partitioning. Random Forest [59] is an ensemble learning method for classification, regression, and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set, offering more accuracy through averaging the results of numerous trees. These machine learning models will be used to assess how well synthetic data can be used to replace real data for the training of predictive models. The next section talks about the methods employed to answer the research questions discussed earlier.

Chapter 3

Methodology

3.1 Data Collection and Preprocessing

3.1.1 Datasets

In this study, we analyzed three datasets: the MIMIC dataset from Physionet [60], the Diabetes Dataset from the UCI Machine Learning Repository [61], and the Adult Income dataset, also from UCI [62]. Table 3.1 outlines the specific characteristics of these datasets. These datasets were selected based on their varying sizes, the number of features they contain, the class ratio of the feature being predicted, and their incorporation of both categorical and continuous features, offering a diverse range of data types for analysis. This variability across datasets is crucial for validating the hypothesis's generalizability across different data scenarios.

As a preliminary step to employing the synthetic data generation pipeline, thorough preprocessing of the datasets is undertaken to ensure data readiness. This preprocessing involves several steps, universally applicable across all datasets, alongside a set of dataset-specific procedures. Initially, the process entails distinguishing between categorical and continuous variables within each dataset. For continuous variables, an examination for missing values is conducted, leading to the removal of rows where such values are absent in less than 5% of cases. Concurrently, categorical variables undergo transformation through LabelEncoder, facilitating their subsequent analysis.

MIMIC-III stands as a substantial and publicly accessible database that contains de-identified health-related information of more than forty thousand individuals who were

admitted to the critical care units at Beth Israel Deaconess Medical Center during the period from 2001 to 2012 [60]. The Fields of ethnicity, gender, death, religion, marital status, insurance, and age are sampled from MIMIC-III to create a profile for each patient. Additional binary flags for select diagnoses of sepsis, birth, chest pain, hypertension, and overdose are recorded for each patient over all their admissions.

The UCI Diabetes dataset encapsulates a decade of clinical records from 1999 to 2008, sourced from 130 hospitals and integrated delivery networks across the United States [61]. It comprises individual rows that detail the hospitalization records of diabetes patients, including their laboratory tests, medication details, and hospital stays of up to 14 days. In the UCI Diabetes dataset, readmission data is categorized based on the timeframe of patient readmissions: '¡30' denotes readmission within less than 30 days, '¿30' signifies readmission after more than 30 days, and 'No' indicates no record of readmission. For the purposes of analysis, this dataset was preprocessed to consolidate the readmission information into two distinct categories. These categories include patients readmitted within a 30-day period (incorporating the '¡30' classification) and those not readmitted (encompassing both '¿30' and 'No' readmission records). This simplification aids in a clearer analysis of readmission patterns within the patient population studied.

The UCI Adult Income dataset underwent preprocessing that necessitated less specialized techniques. Barry Becker extracted this dataset from the 1994 Census database. It is designed to facilitate the prediction of whether an individual's income surpasses $50,000 per year, based on census data. This dataset is also referred to as the "Census Income" dataset.

Table 3.1: Summary of Datasets

Attribute	MIMIC	Diabetes	Adult Income
Number of Records	58,976	101,766	48,842
Number of Attributes	13	47	14
Primary Attributes	ethnicity, gender, expire_flag, religion, time_in_hospital, age	race, gender, diagnosis, time_in_hospital, readmitted_flag	age, workclass, education, education-num, income_category
Number of Categorical Variables	11	40	9
Number of Continuous Variables	2	7	5
Class Ratio	0 - 61%, 1 - 39%	0 - 88%, 1 - 12%	0 - 75%, 1 - 25%
Source	[60]	[61]	[62]

3.2 Conditional Generative Adversarial Networks (GANs)

The state-of-the-art model under conditional GANs is Conditional Tabular GAN (CTGAN), which is widely used for generating synthetic tabular data for various applications. Among various machine learning algorithms to generate synthetic data, Generative Adversarial Networks (GANs) gained considerable prominence [37]. One of the most popular applications of GAN was the creation of fake images. However, traditional GAN models do not work well with tabular data, such as EHR data, which are mostly tabular data with mixed data types, including continuous and categorical variables. CTGAN is a modified version of the traditional GAN model [43, 63].

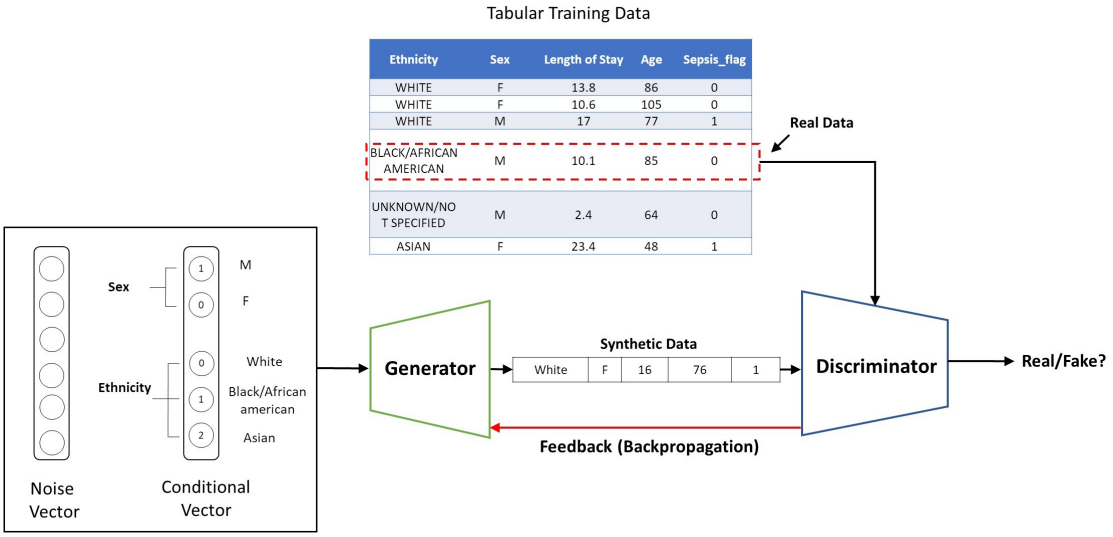


Figure 3.1: Training Process of CTGAN Using MIMIC Dataset as Reference

As illustrated in Figure 3.1, the CTGAN architecture consists of two dense neural networks, the Generator and the Discriminator. The Generator creates a fake/synthetic version of the data, which looks like the real data, while the Discriminator tries to identify if the generated data is fake or real. Although the training data is first preprocessed (for example, categorical variables encoded) before feeding into the pipeline, the CTGAN model performs its own preprocessing before starting the training process. CTGAN has a unique way of handling mixed data types using a mode-specific normalization technique. The continuous variables are normalized to have zero mean and unit variance, whereas the categorical variables are represented in an embedded space [64].

To begin the training process, the model first takes random noise as input and a conditional vector based on the real data containing information about the conditions or constraints under which the synthetic data should be generated. The Generator creates some synthetic samples using this set of information which are sent to the Discriminator along with samples from the real data with the desired columns. The Discriminator then assigns a label to these samples as 'synthetic' and 'real'. The Discriminator sends the learned information back to the Generator using backpropagation. The Generator incorporates this feedback from the Discriminator and tries to improve its performance by creating samples that are fake but close to the real data. During the training process, the Generator aims to create more real-looking synthetic data, and the Discriminator tries to distinguish how to classify the real and synthetic data better. This process continues till the Generator

becomes efficient in creating synthetic samples that the Discriminator classifies as real.

The training process can be controlled using different hyperparameters, such as the number of epochs, the number of iterations/loops of the training process, the learning rate, and the batch size. After each iteration of training, the weights in the neural network within the Discriminator and the Generator get updated.

After completing the training process, the Generator and Discriminator weights get fixed. The next step is called the Generator Forward Pass (Figure 3.2), where the Generator with learned weight takes the random noise and the conditional vector to generate synthetic data. At this stage, the Generator has learned to map this input to a distribution that closely resembles the real data distribution. As illustrated in Figure 3.2, the generated synthetic data retains the same columns as the real data and the same categories for the categorical columns.

The next step in the synthetic data generation process was to check how closely the candidate synthetic data matches with that of the real data (utility and fidelity), while preserving privacy and addressing re-identification risk concerns.

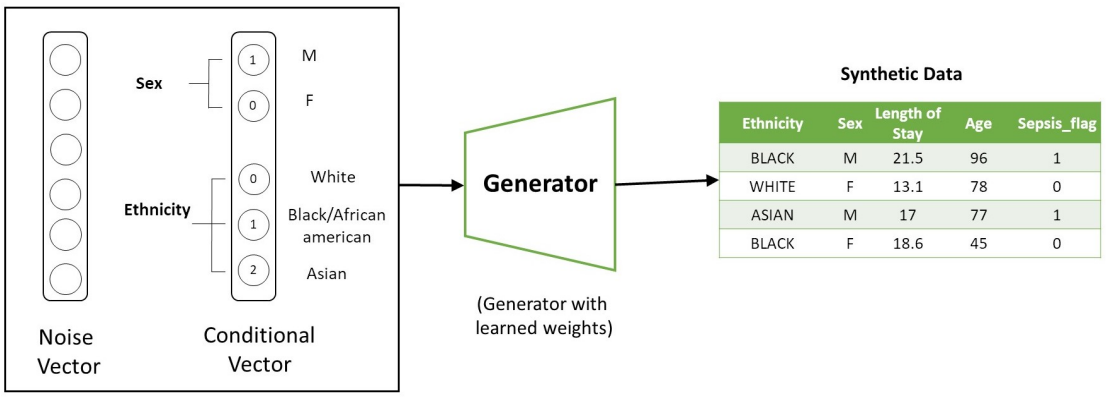


Figure 3.2: Generator Forward Pass of the CTGAN model

3.3 RealTabFormer

In the domain of synthetic data generation, particularly concerning the production of relational datasets, the REaLTabFormer (Realistic Relational and Tabular Transformer) model

represents a significant advancement [65]. This innovative model adeptly manages the complexities associated with simulating relational databases, where it's essential to accurately model not only a "parent" table but also the intricate relationships spanning multiple tables. The REaLTabFormer operates by initially constructing a parent table through an autoregressive mechanism based on the GPT-2 model, followed by the generation of the relational dataset, which is contingent upon the previously created parent table, utilizing a sequence-to-sequence (Seq2Seq) approach. To enhance its functionality and ensure the integrity of the generated data, the model incorporates target masking as a novel regularization technique to mitigate the direct replication of training data. Additionally, it employs the Quantile difference statistic alongside statistical bootstrapping as innovative methods to rigorously assess and prevent overfitting. In real-world dataset evaluations, REaLTabFormer has demonstrated superior ability in capturing relational structures when compared to baseline models, achieving unprecedented performance on predictive tasks across extensive non-relational datasets directly, without the necessity for model fine-tuning. Moreover, in addressing the critical aspect of privacy preservation within synthetic data generation, the model integrates the distance to closest record (DCR) metric and statistical bootstrapping to robustly identify and prevent the model from inadvertently "memorizing" and replicating exact observations from the training data.

3.4 Evaluation Metrics for Synthetic Data

3.4.1 Utility

Utility is a metric used to determine the usefulness of synthetic data for predictive and modelling purposes. The idea is to understand if the generated synthetic data can be used for developing predictive models, in replacement of the real data. It is done by using the real data and synthetic data for training the predictive models, and comparing their performance. Existing literature on the synthetic data domain suggests three main approaches to assessing the utility of synthetic data ([47] and [45]),

(a) Workload-aware evaluation: In this evaluation, various metrics are generated to analyze specific feasibility analyses that are performed on real and synthetic datasets and then compared.

(b) Generic data utility assessment: In this type of assessment, the focus is not on any particular analysis that would be conducted on the synthetic data. Rather, generic structural properties, such as the distance between the real and synthetic data, are evaluated. The metrics generated are usually set to be constrained within a specific range (for

example, 0 to 1), and based on empirical evidence from past studies and existing literature, an interpretation is assigned to the defined range (i.e., does 0 represent high or low? Or, does 0 represent close or distant).

(c) Subjective assessments of data utility: In this type of assessment, domain experts are asked to assess a random mix of real and synthetic records and to classify whether the data is real or synthetic. Standard classification accuracy metrics (such as the F-score or the area under the receiver operating characteristic curve) are used to assess the classification accuracy.

Although these metrics were developed to assess the utility, the three approaches could also be applied to assess the fidelity. For example, the Generic data utility assessment yields scores that indicate how close or far the real and synthetic data are in terms of structure or distribution. These same metrics also give an idea of how closely synthetic data resembles.

Within the workload evaluation, the performance of synthetic data is measured in terms of their ability to be used for predictive machine learning models. This evaluation is termed as Machine Learning Efficiacy (MLE).

Figure 3.3 outlines the procedure for evaluating MLE of synthetic data, aimed at determining the viability of synthetic data for predictive modeling. The process begins with generating synthetic data from real data. Subsequently, two ML models are developed: one trained on real data (M_r) and another on synthetic data (M_s), both employing 5-fold cross-validation to ensure robustness. Each model is then tested on unseen real data to compare their performance. The percentage difference in performance metrics between M_r and M_s quantifies the effectiveness of using synthetic data for training ML models. The percentage difference between the two metrics can be calculated as:

$$\text{Percentage Difference} = \frac{M_r - M_s}{M_r} \times 100\% \tag{3.1}$$

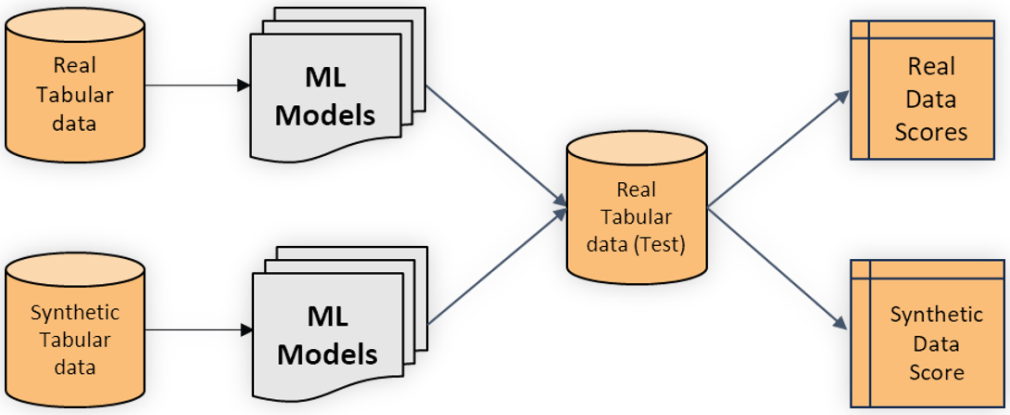


Figure 3.3: Measuring Machine Learning Efficacy of Synthetic Data

For Deep Learning models, we apply hyperparameter tuning using TPE optimizer from the Optuna library, to optimize the results and then calculate the evaluation metrics for the different data scenarios (Figure 3.4). The scenarios considered are: Train on real and test on Real, Train on Synthetic and Test on Synthetic, Train on Synthetic and Test on Real, and Train on balanced synthetic data and test on real data.

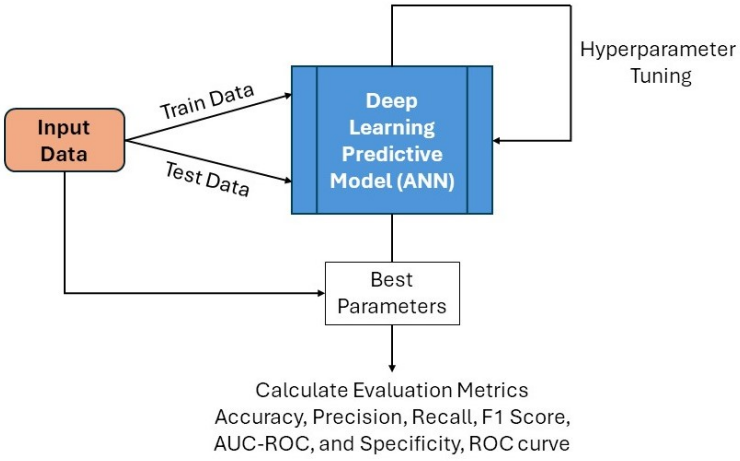


Figure 3.4: Evaluation of Utility of Synthetic Data using Deep Learning model

3.4.2 Fidelity

For assessing the quality of the synthetic data, the statistical distribution of the synthetic and real data is measured. The metric used to compare the distribution is Hellinger Distance (HD). The HD provides a summary statistic, which is a measure of the difference in distribution between each variable in the real (P) and synthetic (Q) datasets. More specifically, it is a probabilistic score that measures between 0 and 1, where 0 indicates no difference in the distribution between real and synthetic datasets [47, 48]. Given two probability distributions $P = \{p_1, p_2, \ldots, p_n\}$ and $Q = \{q_1, q_2, \ldots, q_n\}$, the HD between P and Q is defined as ((Equation 3.2)):

$$H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_i (\sqrt{p_i} - \sqrt{q_i})^2} \tag{3.2}$$

- p_i and q_i are the probabilities of event i in the distributions of P and Q, respectively.
- $\frac{1}{\sqrt{2}}$ is the normalization factor to bind the Hellinger Distance values between 0 and 1.
- n represents the number of distinct events or outcomes for which the probability distributions P and Q are defined.

After calculating HD for each variable for the real and synthetic datasets, we carried out an overall assessment of the HD for all variables, the median and the interquartile range for the real and synthetic data were computed and assessed to check their proximity to 0. A high-utility dataset should have overall HD score closer to 0.

In order to assess the fidelity of the synthetic data, bivariate statistics were computed from both the real and the synthetic datasets, and the absolute differences were computed. If the real and synthetic datasets had high fidelity (i.e., the synthetic dataset closely resembled the real dataset), then the absolute difference would be close to 0 or very small.

For any two continuous variables, the absolute differences between Pearson correlations in the real and synthetic data were evaluated to obtain fidelity in terms of bivariate statistics (Equation 3.3).

$$\Delta XY = \left| \rho_{XY_{\text{Real}}} - \rho_{XY_{\text{Synthetic}}} \right| \tag{3.3}$$

Here, X and Y denote the two continuous variables, whereas ρ_{XY} is the Pearson correlation coefficient for X and Y.

In contrast, for categorical variables, the absolute differences for Chi-square statistics in the real and synthetic data were evaluated as given in Equation 3.4.

$$\Delta\chi^2 = \left|\chi^2_{\text{Real}} - \chi^2_{\text{Synthetic}}\right| \tag{3.4}$$

3.4.3 Privacy Protection

The first level of privacy protection is to remove the directly identifiable variables, such as names, phone numbers or personal IDs, that can be used to directly identify the individual. These variables are chosen from the HIPAA Safe Harbor standards [3]

The second level of privacy protection is to remove the unique combinations of quasi-identifiers. Indirect or quasi-identifiers refer to data elements that, either in isolation or when combined with other indirect or quasi-identifiers or additional information, have the potential to indirectly recognize an individual [66]. It is important to remove the small cell sizes of these quasi-identifiers which can act as outliers and therefore pose a high risk of reidentification. For this work, quasi-identifier combinations with counts less than 5 are removed from the real data, taking reference from Simon et al [67], however, this threshold can be changed based on the stakeholders. The methodology for removing the outliers is mentioned below.

In light of the understanding that it is impossible to entirely eliminate privacy [68] and reidentification risks, this study integrates the privacy and identity disclosure risk assessment framework devised by Khaled et al [47]. This model is instrumental in quantifying the residual risk associated with data anonymization processes. The assessment of risk is bifurcated into two primary categories: the risk of identifying real individuals from synthetic data (Real-to-Synthetic Identification Risk) and the risk of constructing synthetic profiles that can be traced back to real individuals (Synthetic-to-Real Identification Risk) as shown in Equation 3.5. The overarching risk score attributed to the synthetic dataset is determined by considering the higher value between these two risks. Following recommendations from the European Medicines Agency and Health Canada, a predetermined risk threshold, denoted as 0.09 [69], serves as the benchmark for acceptable risk levels in this context.

$$\max\left(\frac{1}{N}\sum_{s=1}^{n}\left(\frac{1}{f_s}\times I_s\times R_s\right), -\frac{1}{n}\sum_{s=1}^{n}\left(\frac{1}{F_s}\times I_s\times R_s\right)\right) \tag{3.5}$$

Where,

- N = The number of records in the real data.
- n = The number of records in the synthetic data.
- s = An index to count records in the synthetic data.
- f_s = The equivalence class group size in the synthetic data for a particular record s in the synthetic data. The equivalence class is defined as the set of records with the same values on the quasi-identifiers.
- F_s = The equivalence class group size in the real data for a particular record s in the real data. The equivalence class is defined as the set of records with the same values as the quasi-identifiers.
- λ'_s = Adjustment to account for errors in matching and a verification rate that is not perfect.
- I_s = A binary indicator of whether records in the real data match a record in the synthetic data.
- R_s = A binary indicator of whether the adversary would learn something new if records in the real data match a record in the synthetic data.

Despite the various measures to ensure privacy and minimize reidentification risk, there is always a residual risk that needs to be ascertained, as the risk can never be minimized to zero [68]. Therefore, the privacy and identity disclosure risk assessment model proposed by Khaled et al. (2020) was adopted to evaluate the privacy and reidentification risk. The risk score under this model could be defined using the equation below [47]. The risk score can be simplified to two parts: Real-to-Synthetic Identification Risk, and Synthetic-to-Real identification Risk. The maximum of both of these risks is taken to be the overall risk of the synthetic dataset. Under the guidance of the European Medicines Agency and Health Canada, an acceptable risk threshold of 0.09 is used. Other approaches for privacy protection include Membership Inference Attack and Attribute Inference Attack, which are not explored in this study but can be studied in future works.

3.5 Hyperparameter Tuning in Synthetic Data Generation

Deep learning (DL) algorithms such as GAN, transformers, and gradient boosting for SDG involve a number of parameters to be set before training. Hyperparameter tuning (HPT) strategies are second-level optimization procedures that try to minimize the expected generalization error of an algorithm over a hyperparameter search space using an objective function [70, 71]. In contrast to model parameters, which are learned during training, these tuning parameters (hyperparameters) have to be carefully selected to optimize model performance. Users have typically 3 choices for selecting an appropriate hyperparameter configuration for a specific dataset: (1) use default hyperparameter values as designed, (2) manually configure hyperparameter values based on recommendations from literature, experience, or trial-and-error, or (3) use HPT strategies [70].

The main goal of HPT is to automatically tune hyperparameters for users to apply machine learning models to practical problems effectively [72, 73]. Although HPT for classification and regression tasks often have a clear choice for objective functions such as any of the metrics computed from the confusion matrix, the choice of the objective function is not so clear for SDG models. As synthetic data is evaluated in a multitude of different ways such as MLE, univariate distribution comparisons, discriminator measures, multivariate correlations, and privacy metrics [74, 75, 76, 65, 77], it is unclear what are the best strategies to tune SDG hyperparameters. Equation 3.6 represents the hyperparameter optimization problem where x^* is the set of hyperparameters of X that minimizes the objective function $f(x)$. It is to be noted that we can maximize the objective function as well, instead of minimizing it. It depends on the nature of the objective function.

$$x^* = \arg\min_{x \in X} f(x) \tag{3.6}$$

Recent literature states the importance of hyperparameters on the performance of SDG models but there still lacks a clear framework for HPT [75, 76, 78]. In addition, it is equally important to have a HPT strategy that can be efficiently applied. Although machine learning efficacy is an important metric for SDG models, it can be expensive to compute as an objective function in a multi-objective HPT framework. To tackle the problem of inefficiencies of MLE as an HPT objective function, we propose to use differential pairwise correlation (DPC) as an alternative objective to MLE, aiming to reduce computational costs.

For HPT of CTGAN, the hyperparameters considered are batch size, generator learning rate, and discriminator learning rate, generator decay, discriminator decay, generator dimension, discriminator dimension, epochs. The batch size is crucial as it determines the number of data samples processed in a single training step, affecting the model's learning dynamics and memory requirements. The learning rates for both the generator and discriminator, denoted as generator_lr and discriminator_lr respectively, are pivotal in determining the speed at which these components of the GAN learn and adapt. These rates dictate the size of the steps taken in the optimization process, where smaller steps can lead to more precise convergence, albeit at a potentially slower rate.

Equally important are the decay rates for the generator and discriminator, referred to as generator_decay and discriminator_decay. These parameters are integral to the Adam optimizer, introducing a regularization aspect that helps in curbing overfitting by penalizing larger weights. This is crucial in maintaining a balance in the learning process, ensuring that neither the generator nor the discriminator becomes too dominant.

The dimensions of the generator and discriminator, termed as generator_dim and discriminator_dim, define the size of the output samples for each of their respective layers. These dimensions are not just numbers; they significantly influence the capacity of the generator to create diverse and complex samples and the ability of the discriminator to accurately classify these samples.

Lastly, the number of epochs plays a vital role. This parameter sets the total number of complete passes the model makes through the entire training dataset. The right number of epochs is a delicate balance – too few might result in underfitting, while too many might lead to the model memorizing the training data, known as overfitting. Each of these parameters, when finely tuned, contributes to a harmonious balance in the CTGAN model, ensuring an effective and efficient training process.

Three choices are provided for each of the three hyperparameters chosen: batch size can be one of 50, 100, or 200, generator learning rate can be one of 1e-3, 1e-4, or 1e-5, and discriminator learning rate can be one of 1e-3, 1e-4, or 1e-5. These hyperparameter values result in a grid of 3x3x3 with 27 unique combinations. As there are 27 unique combinations, under grid search, a total of 27 corresponding trials are run.

Grid search is the conventional method of hyperparameter optimization, where the model is trained across all combinations of all hyperparameters[79]. The method forms a grid of all the hyperparameters and their values and then creates unique combinations of these hyperparameters. For each trial of optimization, the aim is to find an optimal value of the objective function. The objective functions are discussed in the next section, which can be minimized or maximized based on the nature of the function. The gird search

algorithm is easy to implement but is computationally expensive considering the number of hyperparameter combinations considered for each trial. Additionally, with the increasing number of hyperparameters and the search space, the computational cost increases exponentially.

In contrast to Grid search, **Random search** [80] algorithm does not consider all combinations of hyperparameters. Instead, a fixed number of combinations are used for the set of trials. The number of combinations to be considered is chosen randomly. First, a search space is defined, and points are sampled randomly within the search space. In case of Random search, the search space is defined in a continuous distribution within the specified range to take advantage of the random sampling of points. This process is less computationally expensive than the Grid Search because it does not consider all the hyperparameters. On the other hand, there is a chance that some hyperparameter combinations might be missed in the random search method.

Another state-of-the-art algorithm for hyperparameter tuning is probability-based optimization models, which include Bayesian Optimization models such as Gaussian Processes (GP), Sequential Model-Based Optimization (SMBO), and notably, the Tree-structured Parzen Estimator (TPE). TPE [81] stands out by utilizing a non-parametric approach, modeling the search space using conditional probabilities to intelligently navigate towards optimal hyperparameter settings. This strategy is particularly adept at handling complex and high-dimensional spaces where traditional methods falter due to computational constraints or the curse of dimensionality. For this research work we use a combination of Random search and TPE search method for HPT.

As discussed above, there are primarily three ways to measure the quality of synthetic data, utility, fidelity and privacy. To generate the best quality of synthetic data, it is important to choose the best objective function based on these quality metrics which will be minimized or maximized (depending on the nature of the objective function) during the HPT process. Utility-based objective functions (MLE) are generally considered for HPT of synthetic data generation [82].

We perform experiments to find out if fidelity-based metrics can be used for HPT of synthetic data generation. The benefit of this approach is two-fold. Firstly, Utility-based objective functions require more computational resources, compared to the fidelity-based metrics. Therefore, the total time required for the HPT will be reduced. Secondly, the use of fidelity-based metrics will enable the development of use-case agnostic synthetic data. In real world, the synthetic tabular data maybe used for more than one use case, and therefore using fidelity based metrics might be a better approach to create synthetic data that can be used for multiple use cases. To validate this, we perform a correlation between

the fidelity and utility metrics for different hyperparameter combinations. Hyperparameter tuning of the synthetic data generation is performed using fidelity metrics (HD, and DPC) as the objective function. During HPT, for each hyperparameter combination, the utility metric is calculated and stored.

Since there are two fidelity-based metrics: Fidelity - Hellinger Distance (uni-variate measure) and Differential Pairwise Correlation (bi-variate measure), we perform multi-objective optimization to get the Pareto front that gives the best fidelity measurement.

For this thesis, the computational work was carried out on a system powered by an AMD EPYC 7543 32-Core Processor and supported by 512 GB of RAM, ensuring efficient processing for extensive datasets. The system has an NVIDIA A30 GPU with 24576MiB capacity, running on Windows OS with an AMD64 architecture. This setup, equipped with NVIDIA driver version 537.70 and CUDA Version 12.2, provided a robust platform for advanced computational tasks and deep learning algorithms essential for the research.

For the hyperparameter tuning of the synthetic data generation, **Stratified Random Sampling** technique is used which is a method to develop a subset dataset that represents the entire dataset[83]. This method is important considering the size of larger datasets in the real world. For high-dimensional datasets, the hyperparameter tuning of the synthetic generation method is computationally expensive, since, for each combination of hyper-parameters, the synthetic generation will take place for the entire dataset, followed by the calculation of evaluation metrics. Therefore, a method like stratified random sampling helps save computational resources and time and makes the hyperparameter tuning method more efficient. The algorithm for random sampling is shown in Algorithm 1.

Algorithm 1 Create stratified sample based on one variable

Initialize *stratified_sample* as an empty DataFrame
Define *n_samples* as per requirements
Calculate *strata_proportions* as the normalized value counts of 'variable' in *real_data*
Determine *samples_per_stratum* by rounding up the product of *strata_proportions* and *n_samples* to the nearest integer
for each stratum, *n* in *samples_per_stratum*.items() **do**
    Filter the original DataFrame *df* for the current stratum
    stratum_sample ← *df[df['variable'] == stratum].sample(n, random_state=1)*
    stratified_sample ← pd.concat([*stratified_sample*, *stratum_sample*])
end for
return *stratified_sample*

3.6 Hyperparameter Tuning of Predictive Models Using Synthetic Data

When building predictive models using health data that contains sensitive information about individuals, the training data may not leave the premises of data sources. However, these health organizations may lack the computational resources to train and optimize the performance of these predictive models.

The hypothesis of the research posits that HPT in predictive modelling can be achieved by leveraging synthetic data. If proven, this methodology could circumvent the challenges associated with the inaccessibility of real data for external analysis. Synthetic data could offer a viable solution for resource-intensive HPT of predictive models using external computing resources of an organization, as illustrated in 3.5.

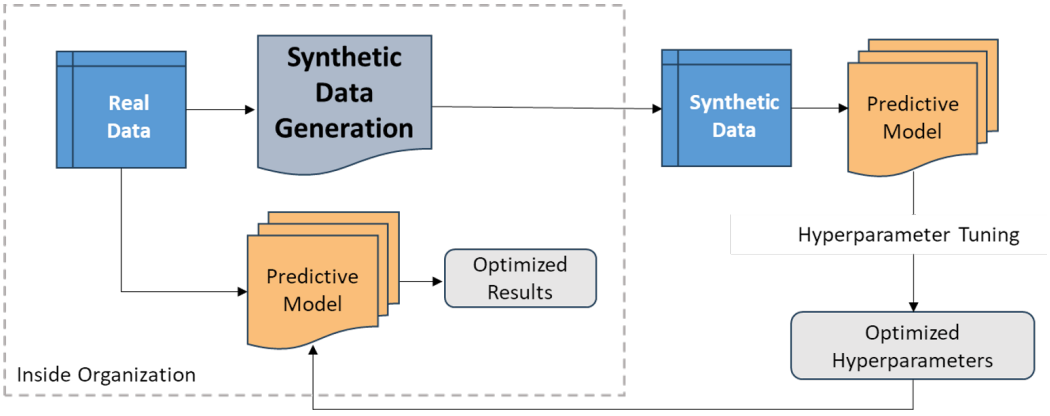


Figure 3.5: Hyperparameter tuning of predictive model using Synthetic Data

Therefore, in this study, we use synthetic data for HPT of the predictive model, and then use the best hyperparameters for the predictive model using real data.

Let M be a machine learning model parameterized by a set of hyperparameters $\theta \in \Theta$, where Θ is the hyperparameter space. The model's performance is evaluated by a function f, such that $f(M, D) = p$, where p is the performance metric (e.g., accuracy, F1 score) of the model M with hyperparameters θ on dataset D.

Hyperparameter tuning involves finding the optimal set of hyperparameters θ^* that maximizes (or minimizes) the performance metric on the synthetic dataset D':

$$\theta^* = \arg\max_{\theta \in \Theta} f(M_{\theta^*}, D') \tag{3.7}$$

The selected hyperparameters θ^* are then used to train the model on the real dataset D, and the performance is evaluated:

$$p_{\text{real}} = f(M_{\theta^*}, D) \tag{3.8}$$

The correlation between the metrics obtained from the synthetic data tuning and the real data application can be evaluated to validate the effectiveness of the synthetic tuning process. Let p_{synth} be the performance metric on the synthetic data, and p_{real} be the metric on the real data:

$$\text{Correlation} = \rho(p_{\text{synth}}, p_{\text{real}}) \tag{3.9}$$

where ρ denotes a correlation coefficient (e.g., Pearson, Spearman). For our study we have used Pearson Correlation coefficient.

Utilizing synthetic data for hyperparameter tuning before applying these parameters to models trained on real data has significant implications for data security and model optimization. It enables organizations to exploit external computational resources and expertise for model enhancement while adhering to data protection regulations. This approach is especially beneficial in industries dealing with highly sensitive information, such as healthcare or finance, where data sharing is heavily regulated. By validating the effectiveness of hyperparameter settings with synthetic data, researchers can enhance the predictive performance of their models on real datasets, thereby bridging the gap between data privacy concerns and the need for advanced analytical techniques.

3.7 Explainability Using Synthetic Data

Explainability in machine learning refers to the ability to understand and interpret how predictive models make decisions [84]. Explainability is an important consideration when using predictive models for applications that have high-stakes example finance and healthcare. White box models like linear regression, and random forest, are explainable at some levelS. However, it is difficult to explain the predictions obtained using black box models like deep learning models including neural networks. Explainable AI (XAI) aims to make

the operations of AI systems transparent and comprehensible to human users, shedding light on how models process input data to make predictions or decisions.

This study aims to explore the consistency of feature importance across models trained on real and synthetic datasets. The hypothesis posits that there exists a strong correlation between the mean SHAP (SHapley Additive exPlanations) values of variables in models trained on real data and their counterparts in models trained on synthetically generated data. SHAP values provide a robust, game-theory-based method for quantifying the contribution of each feature to model predictions. By comparing these values across real and synthetic datasets, we aim to assess the viability of using synthetic data for model explanation and validation.

Similar to the approach used in the utility, to validate the hypothesis on explainability we train a prediction model on real data and test it on real data, then we train a model on synthetic data, and test it using real data. In both cases the SHAP values of all the feature variables, for all the predictions are calculated. Thereafter, the mean of the SHAP values for each feature variable is taken for real and synthetic data. They are plotted together on one graph to study their correlation. The second approach is to explore the use of synthetic data to explain the models trained and tested on real data [85]. In scenarios where data cannot be released to the public, however the pre-trained models are shared with other organizations, it is difficult to prove explainability of the synthetic data.

Chapter 4

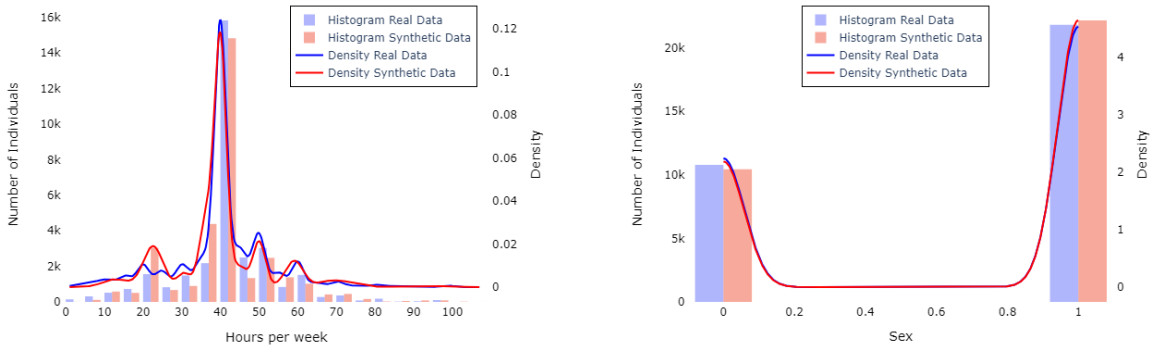
Results and Discussion

4.1 Fidelity Metrics - Representativeness

The synthetic data should statistically represent the real data and its characteristics. To evaluate the ability of the synthetic data to accurately represent the statistics, we compare the frequency and probability distribution of different features from the synthetic and real data. In addition to the fidelity metrics to evaluate the representativeness, we also compare the cohort characteristics of the real and synthetic data.

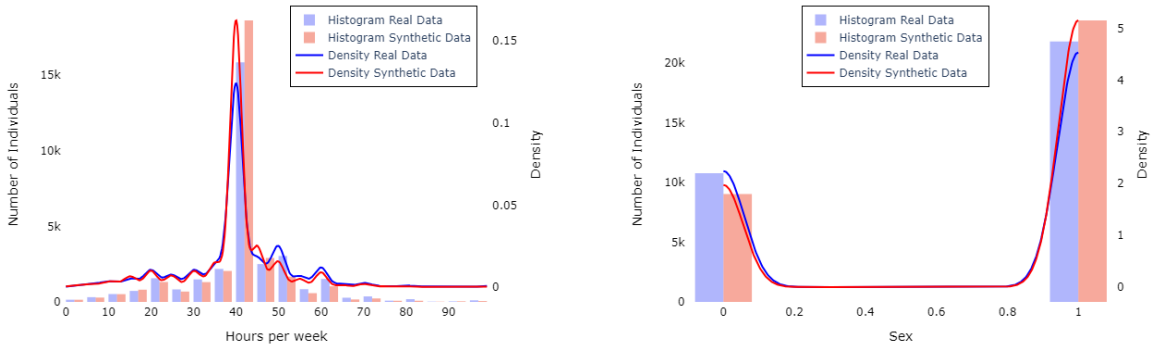
4.1.1 Adult Income Dataset

In evaluating the statistical representativeness of synthetic data, a critical assessment involves contrasting the distribution—both in terms of frequency and density—of features between the synthetic and real datasets. This comparative analysis is illustrated in Figure 4.1 and Figure 4.2, which delineate the distribution patterns of a continuous variable ('Hours-per-week') and a categorical variable ('Sex') generated by the CTGAN and RealTabFormer respectively. These figures elucidate that the distributions of both variables from the CTGAN and RealTabFormer models exhibit a remarkable similarity to those of the original data.



(a) Hours per week (b) Sex

Figure 4.1: Distribution (Frequency and Probability) for 'Hours per week' and 'Sex' variable for real and synthetic data for Adult Income dataset from CTGAN



(a) Hours per week (b) Sex

Figure 4.2: Distribution (Frequency and Probability) for 'Hours per week' and 'Sex' variable for real and synthetic data for Adult Income dataset from RealTabFormer

In the evaluation of synthetic data fidelity, both univariate and bivariate metrics were employed, with Hellinger Distance (HD) utilized for univariate assessment and Differential

Pairwise correlations for bivariate analysis. This dual approach facilitated a detailed examination of the synthetic data's quality, especially in the context of the Adult Income dataset. As illustrated in Figure 4.3, and 4.4, the analysis yielded significant insights. Specifically, features such as *sex* and *education* recorded the lowest HD values, indicating high fidelity in these aspects of the synthetic data. On the other hand, continuous variables like *hours-per-week* showed the highest HD, revealing a potential challenge for the model in capturing extreme values accurately. Nonetheless, the HD values across all features were found to be within an acceptable range, affirming the overall quality of the synthetic data generated.

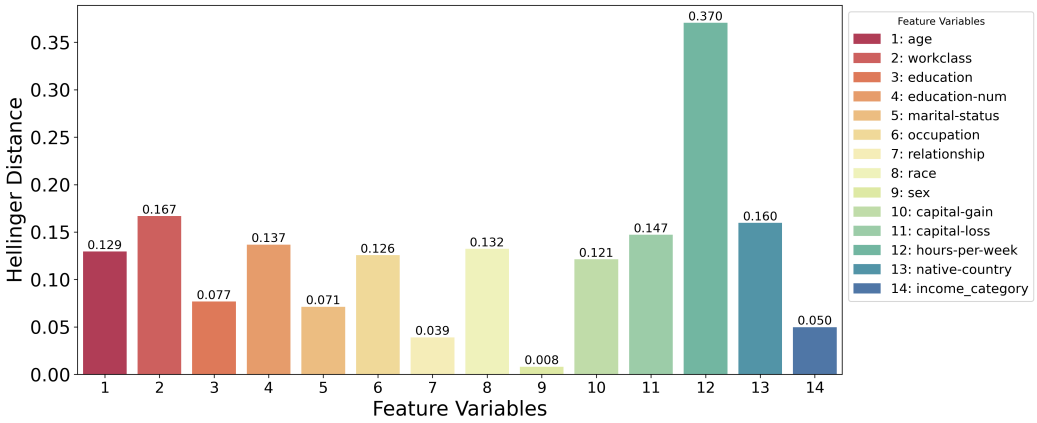


Figure 4.3: Hellinger Distance for Adult Income Dataset - CTGAN

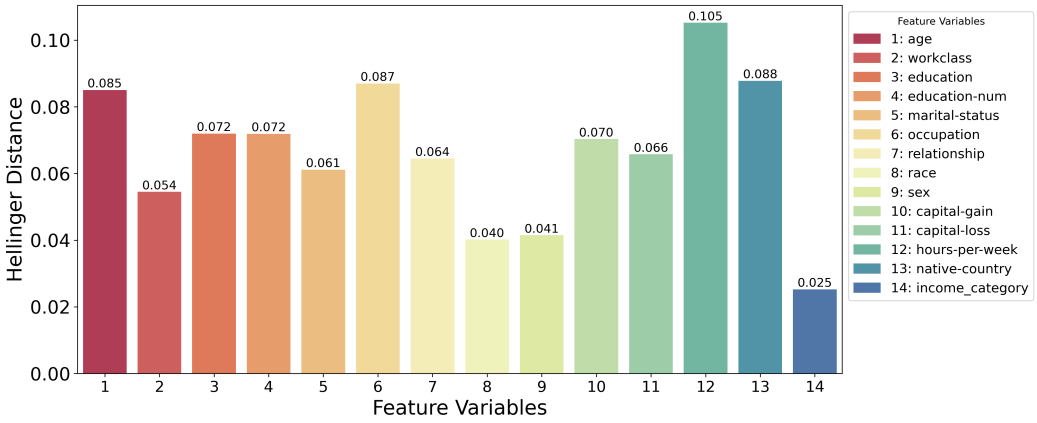
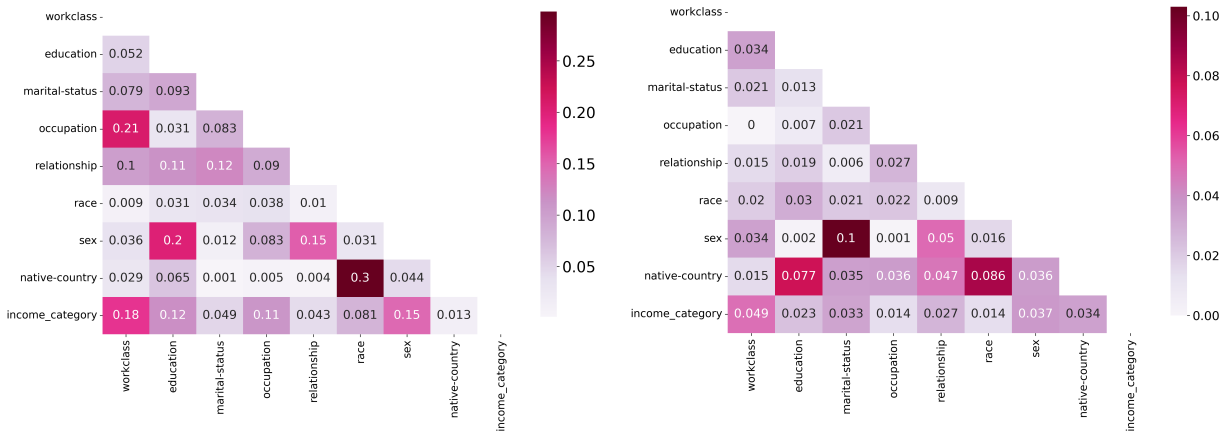


Figure 4.4: Hellinger Distance for Adult Income Dataset - RealTabFormer

Figure 4.5 presents a comparative analysis of differential pairwise correlations for the Adult Income dataset, contrasting synthetic data generated by CTGAN and RealTabFormer methods against real-world data. The visualizations are displayed in the form of matrices, where each cell represents the difference in correlation between the synthetic and real data for each pair of variables. In the CTGAN matrix (a), higher discrepancies are observed in pairs involving the '*occupation*' and '*income_category*' variables, with some notable differences also seen with '*workclass*' and '*sex*'. In contrast, the RealTabFormer matrix (b) exhibits relatively smaller differences across the board, suggesting a closer approximation to the real data's correlation structure. Noteworthy is the '*native-country*' variable, which shows a more consistent correlation pattern with other variables in the RealTabFormer method compared to CTGAN. Overall, the RealTabFormer approach appears to yield a synthetic dataset that more closely mirrors the correlation characteristics of the original data, indicating its potential superiority in maintaining the statistical properties of the dataset for the analyzed variables.



(a) CTGAN (b) RealTabFormer

Figure 4.5: Differential Pairwise Correlation (DPC) Heatmap for real and synthetic data for Adult Income dataset from CTGAN and RealTabFormer

Continuing from the examination of fidelity metrics, the comparative analysis of predicted class counts between real and synthetic data offers further insights, pivotal for applications in binary classification models. As detailed in Table 4.1, this comparison reveals notable variances in class distribution between the original dataset and its synthetic counterparts generated by CTGAN and RealTabFormer.

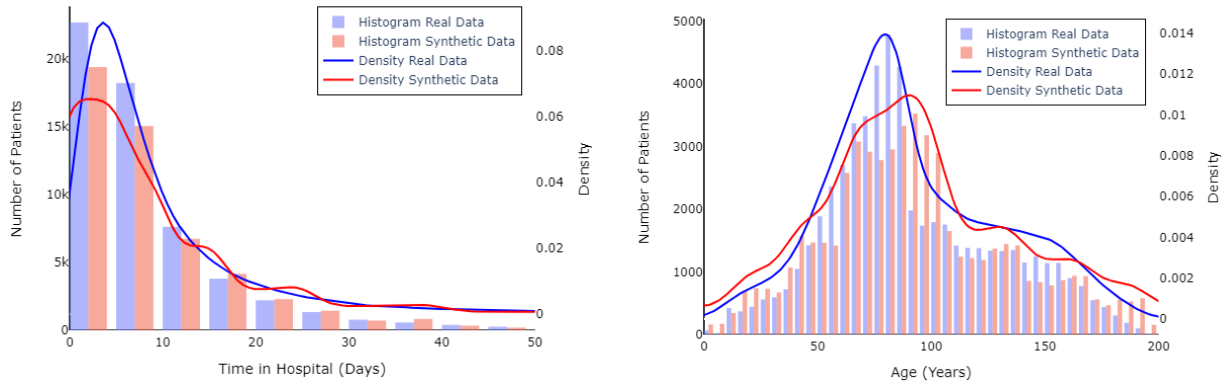
For Class 0 (Income category ≤50K),, which represents the majority class in the Adult Income dataset, the real data count stands at 24,720. The CTGAN model slightly underrepresents this class, generating 22,681 instances, marking an 8.2% decrease. In contrast, RealTabFormer overestimates Class 0 with 25,692 instances, an increase of 3.93%. This discrepancy indicates a variance in each model's ability to replicate the majority class's distribution accurately. On the other hand, Class 1 (Income category >50K), the minority class, is represented by 7,841 instances in the real data. CTGAN significantly overestimates this class with 9,880 instances, resulting in a 26% increase. Conversely, RealTabFormer underestimates Class 1 with 6,869 instances, showing a decrease of 12.39%. This variance highlights each model's differing capacity to capture the minority class's nuances.

Table 4.1: Class Counts for Real and Synthetic Data - Adult Income Dataset

Class Counts - Adult Dataset			
	Real Data	**Synthetic Data**	
		CTGAN	**RealTabFormer**
Class 0	24720	22681 (-8.2%)	25692 (3.93%)
Class 1	7841	9880 (26%)	6869 (-12.39%)
Total	32561	32561	32561

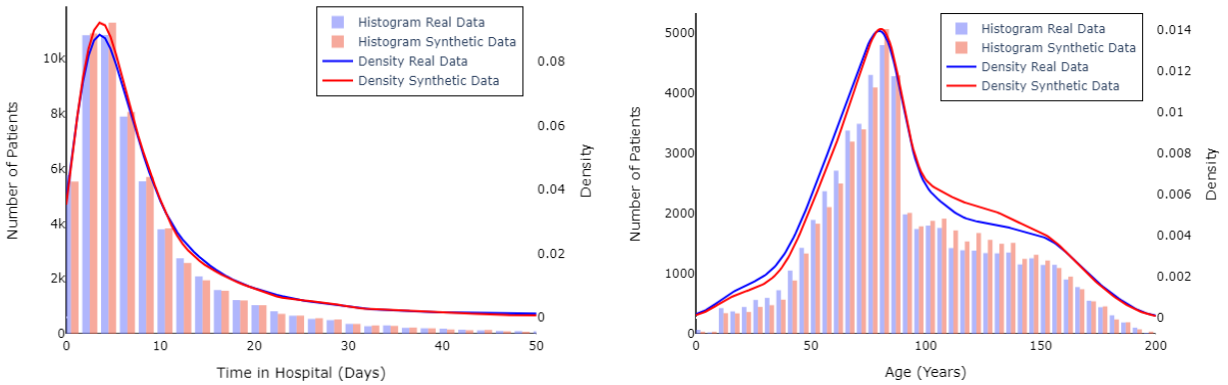
4.1.2 MIMIC Dataset

The distributions of the continuous variable *'Length of Stay'* and the categorical variable *'Age'* from the CTGAN and RealTabFormer, as illustrated in Figures 4.6 and 4.7, exhibit a high resemblance to those in the original MIMIC dataset. This alignment in frequency and density indicates the synthetic data's statistical representativeness. Both models' capacity to accurately mimic the complex distribution patterns of *'Length of Stay'* and *'Age'* highlights their effectiveness in generating synthetic datasets. Such datasets not only preserve the essential statistical properties of real healthcare data but also serve as a reliable basis for further analysis and modeling in the context of sensitive healthcare research. The HD for each feature is shown in Figure 4.8 and 4.9, with the minimum HD value for 'EXPIRE_FLAG', and the maximum HD value for 'Age'.



(a) Time (b) Age

Figure 4.6: Distribution (Frequency and Probability) for 'Length of Stay' and 'Age' variable for real and synthetic data for MIMIC dataset from CTGAN



(a) Time (b) Age

Figure 4.7: Distribution (Frequency and Probability) for 'Length of Stay' and 'Age' for real and synthetic data for MIMIC dataset from RealTabFormer

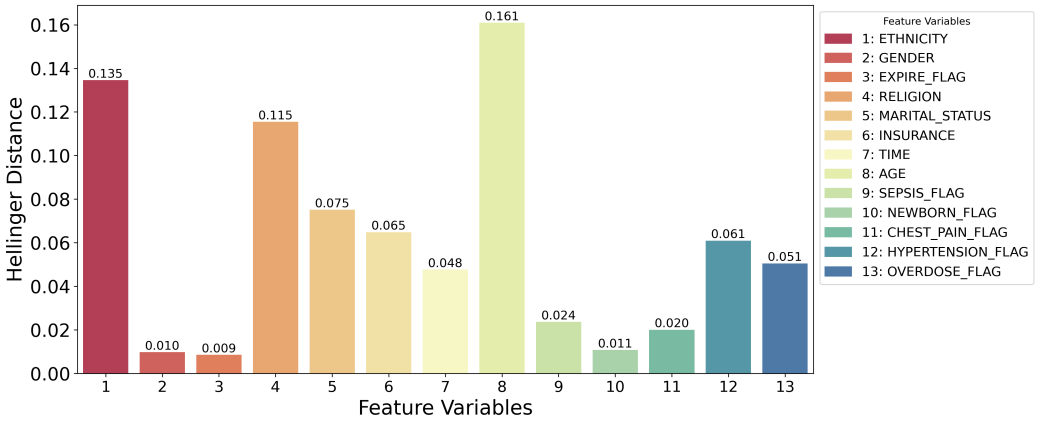


Figure 4.8: Hellinger Distance for MIMIC Dataset

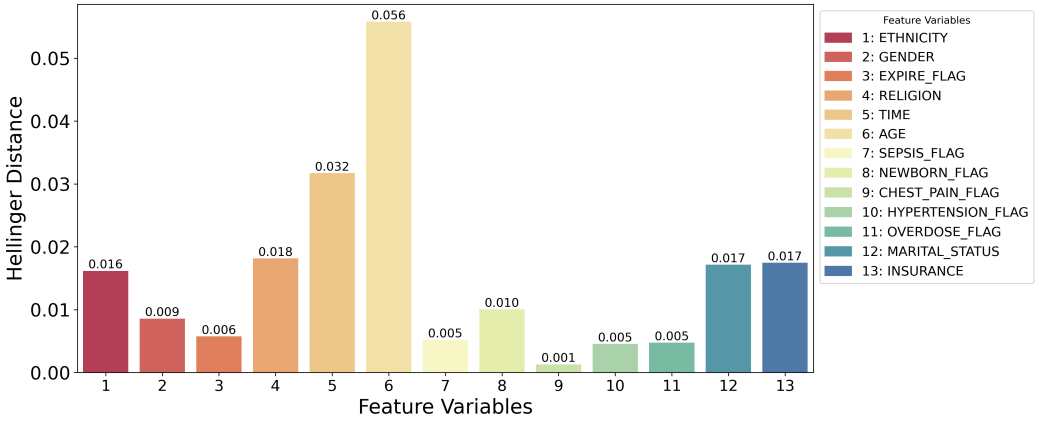
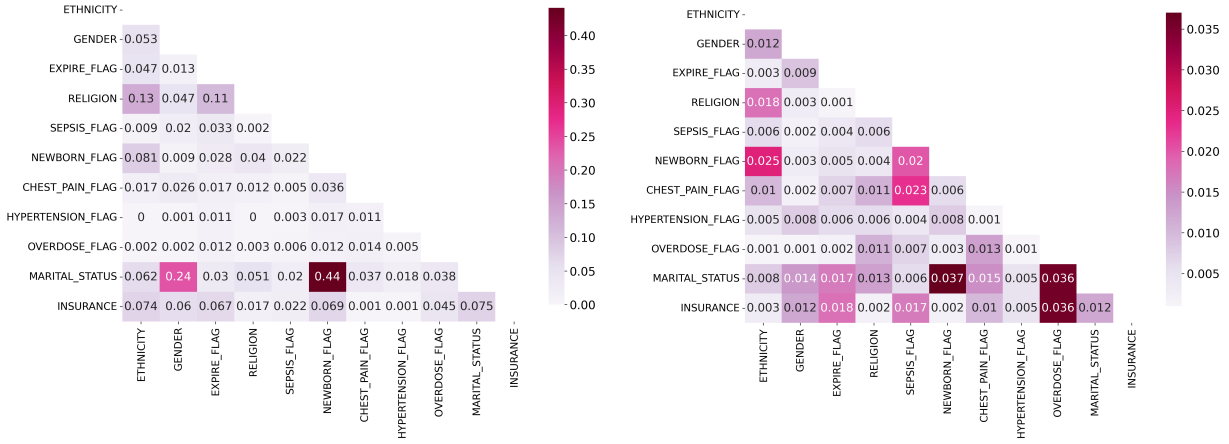


Figure 4.9: Hellinger Distance for MIMIC Dataset - RealTabFormer

Figure 4.10 illustrates the differential pairwise correlation matrices for the MIMIC dataset, contrasting the synthetic data generated by CTGAN (a) and RealTabFormer (b) against the original data. The matrices show the correlation difference for each variable pairing, with color intensities reflecting the magnitude of the differential. The CTGAN matrix indicates more significant variances, especially in pairs involving '*MARITAL_STATUS*' and '*INSURANCE*', where the correlations in synthetic data differ substantially from those in the real dataset. Conversely, the RealTabFormer matrix demonstrates tighter correlation differentials, with most variable pairs showing marginal discrepancies. Notably, '*NEWBORN_FLAG*' and '*CHEST_PAIN_FLAG*' exhibit relatively closer correlations to

the real data when generated by RealTabFormer. This suggests that the RealTabFormer method may be more adept at replicating the underlying statistical relationships within the MIMIC dataset, potentially making it more suitable for applications that require high fidelity to the original data's correlation structure.



(a) CTGAN (b) RealTabFormer

Figure 4.10: Differential Pairwise Correlation (DPC Heatmap for real and synthetic data for MIMIC dataset from CTGAN and RealTabFormer

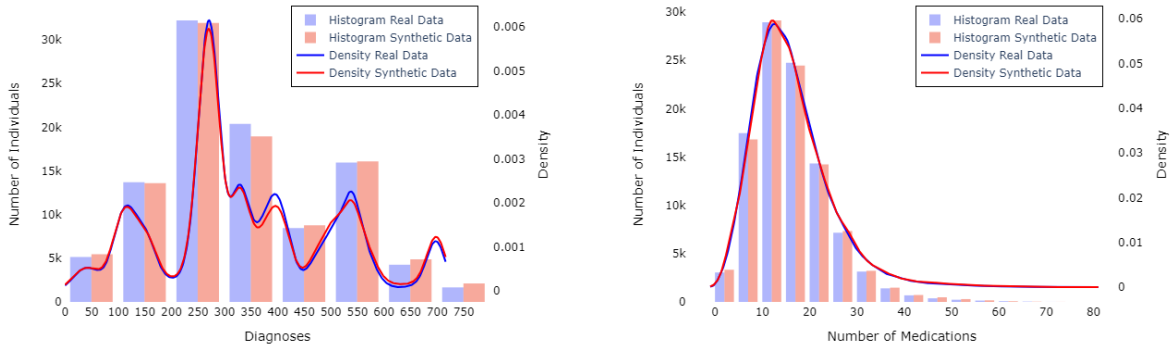
Table 4.2: Class Counts for Real and Synthetic Data - MIMIC Dataset

Class Counts - MIMIC Dataset			
	Real Data	**Synthetic Data**	
		CTGAN	**RealTabFormer**
Class 0	35924	37079 (1.89%)	36390 (1.28%)
Class 1	23052	21897 (-3.05%)	22586 (2.06%)
Total	58976	58976	58976

4.1.3 Diabetes Dataset

The distributions of the continuous variable *'Number of medications'* and the categorical variable *'Diagnosis 1'* from the CTGAN and RealTabFormer, as demonstrated in Figures

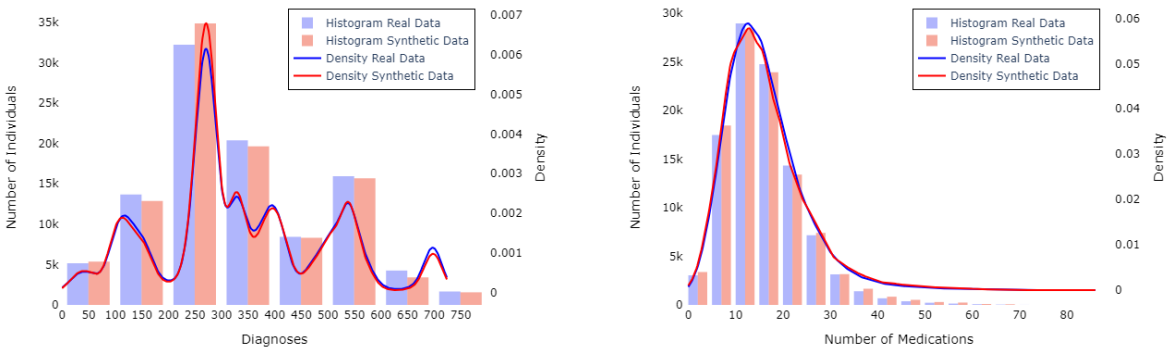
4.11 and 4.12, closely match those in the original UCI Diabetes dataset. This congruence in both frequency and density underscores the synthetic data's accuracy in reflecting the dataset's statistical properties. The ability of the CTGAN and RealTabFormer models to replicate the intricate distribution patterns of *'Number of medications'* and *'Diagnosis 1'* attests to their efficiency in generating synthetic datasets that maintain critical statistical characteristics of the real diabetes data, thus providing a solid foundation for subsequent data analyses and modeling efforts in diabetes research. Additionally, the HDs for Diabetes dataset as shown in Figure 4.13 and 4.14, show small values of HD for all variables, with exceptionally low values for 'readmitted_flag' feature.



(a) Diagnoses type

(b) Number of Medications

Figure 4.11: Distribution (Frequency and Probability) for 'Diagnosis 1' and 'Number of medications' variable for real and synthetic data for Diabetes dataset from CTGAN



(a) Diagnoses type

(b) Number of Medications

Figure 4.12: Distribution (Frequency and Probability) for 'Diagnosis 1' and 'Number of medications' variable for real and synthetic data for Diabetes dataset from RealTabFormer

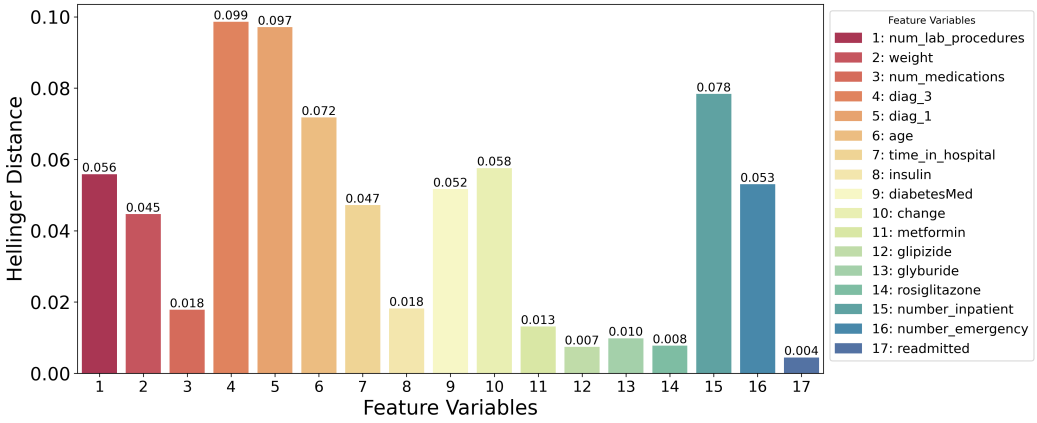


Figure 4.13: Hellinger Distance for Diabetes Dataset

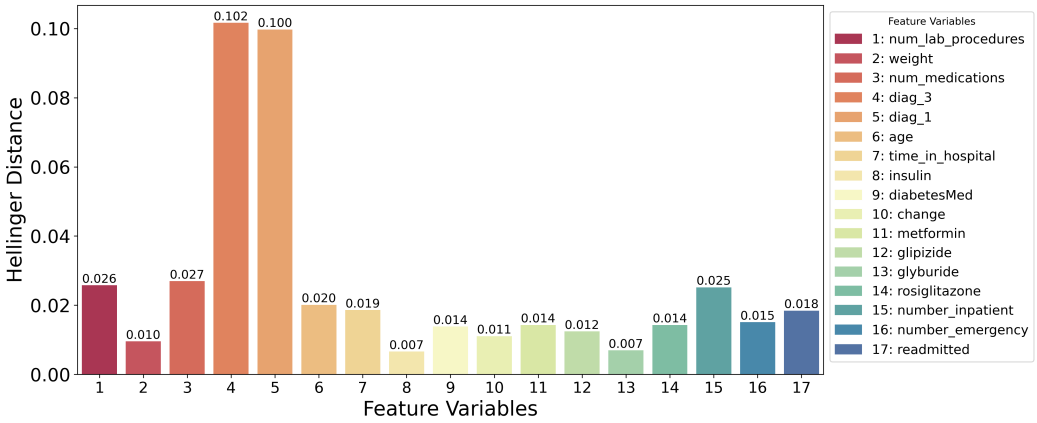
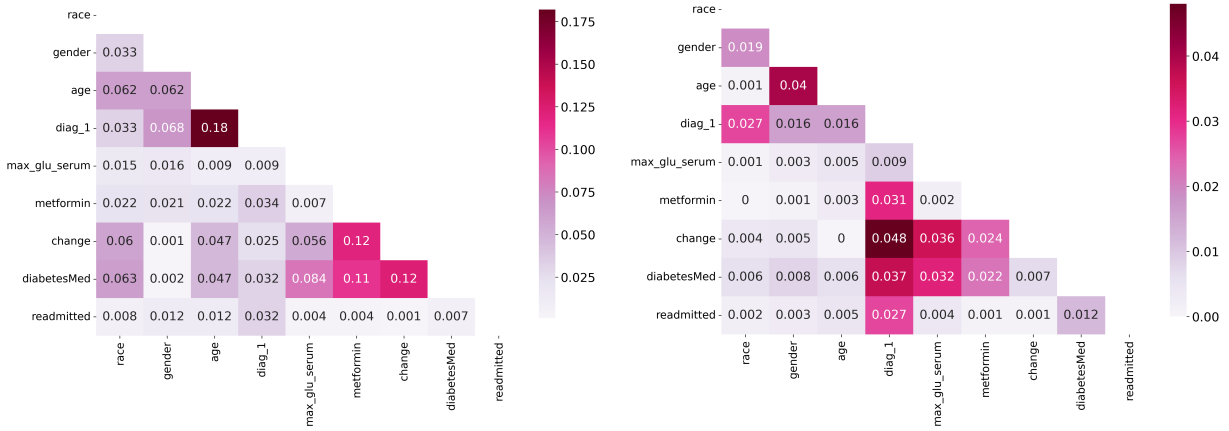


Figure 4.14: Hellinger Distance for Diabetes Dataset - RealTabFormer



(a) CTGAN (b) RealTabFormer

Figure 4.15: Differential Pairwise Correlation Heatmap for real and synthetic data for Diabetes dataset from CTGAN and RealTabFormer

Figure 4.15 shows the differential pairwise correlation of the variables between synthetic and real data. The heatmap comparison between the CTGAN (a) and RealTabFormer (b) models shows that each model has a different pattern of pairwise correlation values among the variables. This suggests that the underlying algorithms in each model capture and represent the inter-variable relationships in distinct ways. Both models demonstrate some challenges in accurately replicating the correlation between '*diag_1*' and other variables.

In CTGAN, there is a notable differential correlation with '*change*' and '*diabetesMed*', whereas in RealTabFormer, the largest discrepancy is again with '*change*', but to a lesser extent. The differential pairwise correlation represents the difference in the Pearson correlation coefficients between the corresponding variables in the synthetic and real datasets. The differential correlations for '*race*' and '*max_glu_serum*' are relatively low across both models, suggesting that the synthetic data generated by CTGAN and RealTabFormer adequately captures the correlation structure between these variables and others in the real dataset. On the contrary, the heatmap for RealTabFormer shows that the most substantial differential correlations do not exceed 0.05, with most values clustered around 0 or below 0.03. This implies a closer alignment with the real data's correlation structure, indicating that RealTabFormer may be more effective in maintaining the statistical properties of this dataset.

The comparison of class distributions between real and synthetic datasets, as presented in Table 4.3, offers valuable insights for binary classification model development within the MIMIC Dataset context. The CTGAN model closely mirrors the real data with slight variances, showing a minimal decrease in Class 0 counts and a moderate increase for Class 1. Conversely, RealTabFormer demonstrates a slight overestimation for Class 0 but a notable underrepresentation of Class 1. These differences underscore the importance of selecting an appropriate synthetic data generation method that aligns with the specific goals of the classification task, emphasizing the potential and challenges of employing synthetic data for model training.

Table 4.3: Class Counts for Real and Synthetic Data - MIMIC Dataset

Class Counts - Diabetes Dataset			
	Real Data	**Synthetic Data**	
		CTGAN	**RealTabFormer**
Class 0	90409	90002 (-0.45%)	92004 (1.76%)
Class 1	11357	11764 (3.58%)	9740 (-14.24%)
Total	101766	101766	101766

The boxplot in Figure 4.16 compares the Hellinger Distance (HD) across three distinct datasets: Diabetes, MIMIC, and Adult Income. The Diabetes dataset shows a compact interquartile range (IQR) with a median HD value near 0.05, indicating a relatively tight

clustering of distances. The MIMIC dataset's boxplot displays a slightly higher median HD than the Diabetes dataset, along with a larger IQR, suggesting greater variability in the HD measurements. The Adult Income dataset presents the widest IQR and the highest median HD, close to 0.10, reflecting the most significant disparity among the three. For the Diabetes and MIMIC datasets, the HD values are largely consistent, as evidenced by the few outliers.

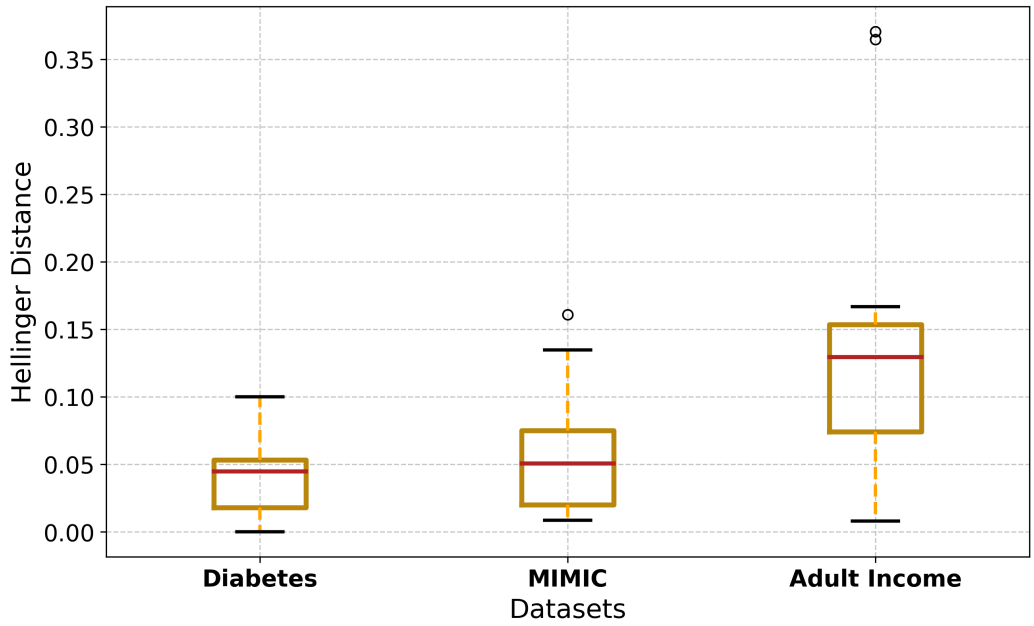


Figure 4.16: Hellinger Distance for Adult Income, Diabetes, MIMIC Dataset

4.1.4 Comparative Analysis of Fidelity Metrics

The fidelity of synthetic data generation was evaluated for CTGAN and RealTabFormer across Adult Income, MIMIC, and Diabetes datasets using Mean HD and Mean Differential Pairwise Correlation (DPC) as summarized in Table 4.4. RealTabFormer outperformed CTGAN in all cases, showing lower Mean HD and Mean DPC values. For instance, in the Adult Income dataset, RealTabFormer reduced the Mean HD from 0.139 (CTGAN) to 0.084 and the Mean DPC from 0.077 to 0.029. This pattern of superior performance by RealTabFormer was consistent across the MIMIC and Diabetes datasets, underscoring its effectiveness in producing higher-fidelity synthetic data.

Table 4.4: Fidelity Comparison of Datasets and SDG Models

	Adult Income		MIMIC		Diabetes	
Metric	**CTGAN**	**RealTabFormer**	**CTGAN**	**RealTabFormer**	**CTGAN**	**RealTabFormer**
Mean HD	0.139	0.084	0.062	0.014	0.041	0.010
Mean DPC	0.077	0.029	0.043	0.009	0.031	0.013

4.2 Utility Metrics - Machine Learning Efficacy (MLE)

4.2.1 Adult Income Dataset

For this study, the usefulness of the synthetic data has been evaluated across different machine-learning models. Table 4.5 shows the difference between the performance of ML models on synthetic and real data. The XGBoost model shows the most significant decline in Recall (-26.58%) and F1 Score (-15.49%), indicating a particularly reduced ability to correctly identify positive cases and a decrease in the harmonic mean of precision and recall, respectively, when trained on synthetic data. Logistic Regression shows an increase in Precision (+10.87%) when using synthetic data. This is an anomaly as precision improves, suggesting that for this model, synthetic data might be leading to more conservative predictions that are more often correct when they predict the positive class. Decision Tree and Random Forest models show a moderate decline across most metrics, with a noticeable decrease in Recall, suggesting a difficulty in identifying all positive instances correctly. The over-representation of Class 1 in synthetic data could be causing models to perform better in predicting positive instances in some cases (e.g., improved precision in Logistic Regression) but also leading to overall performance degradation due to imbalanced training data.

Table 4.5: MLE for all models for Adult Income Dataset

Model	**Accuracy Difference (%)**	**Precision Difference (%)**	**Recall Difference (%)**	**F1 Score Difference (%)**	**Specificity Difference (%)**
XGBoost	0.08 (-9.09%)	0.03 (-4.62%)	0.21 (-26.58%)	0.11 (-15.49%)	0.09 (-9.47%)
Logistic Regression	0.02 (-2.38%)	-0.05 (10.87%)	0.11 (-14.47%)	0.00 (0%)	0.05 (-5.21%)
Decision Tree	0.03 (-3.57%)	-0.06 (13.04%)	0.16 (-20.78%)	0.02 (-3.45%)	0.07 (-7.29%)
Random Forest	0.03 (-3.53%)	0.00 (0%)	0.14 (-17.28%)	0.04 (-6.56%)	0.04 (-4.17%)
MLP	0.02 (-2.35%)	-0.02 (3.7%)	0.10 (-12.99%)	0.03 (-4.69%)	0.04 (-4.21%)
SVM	0.02 (-2.35%)	0.04 (-7.41%)	0.09 (-11.54%)	0.06 (-9.38%)	0.02 (-2.11%)

Table 4.6 shows the performance metrics for different scenarios for training and testing

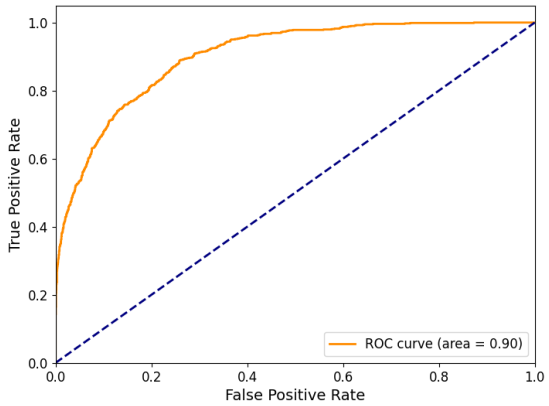
a deep learning model (Feed-forward Neural Network) using real and synthetic data. The first scenario, Train on Real Data, Test on Real Data (No Synthetic Data), provides a baseline performance metric for comparison. High accuracy (84.96%), precision (70.47%), and recall (64.54%) indicate the model performs well on real data.
The model achieves a high AUC_ROC (90.17%) and specificity (91.43%), showing strong capability in distinguishing between classes and correctly identifying negative cases. For Train on Synthetic Data, Test on Synthetic Data (No Real Data), gives slightly lower accuracy (81.12%) and recall (62.25%) compared to training on real data suggest synthetic data might not capture all nuances of the real dataset. Precision is slightly higher (71.76%) than in the real-data scenario, indicating good positive prediction value when the model is kept within the synthetic domain. AUC-ROC (86.53%) and specificity (89.33%) are slightly lower but still strong, indicating the model's effectiveness in class separation and identifying negatives is somewhat diminished but still good in the synthetic domain. Training a model on synthetic data and testing it on real data presents challenges, evidenced by decreased accuracy (75.84%), precision (49.85%), and F1 Score (55.21%). Despite a reasonable recall rate (61.86%), the model's effectiveness in class differentiation and identifying negative cases significantly diminishes, as shown by drops in AUC-ROC (79.01%) and specificity (80.27%).

For the last scenario, training with synthetic data which is balanced (1:1 class ratio), we get an improved accuracy (81.95%), precision (59.40%), and F1 Score (67.80%) compared to training on unbalanced synthetic data suggest that balancing the synthetic dataset helps in generalizing the model to real data.A significant increase in recall (78.95%) indicates that balancing the dataset greatly improves the model's ability to identify positive cases in real data.The AUC-ROC (90.04%) is nearly restored to the baseline of real data training, suggesting excellent overall model performance (Also shown in Figure 4.17).

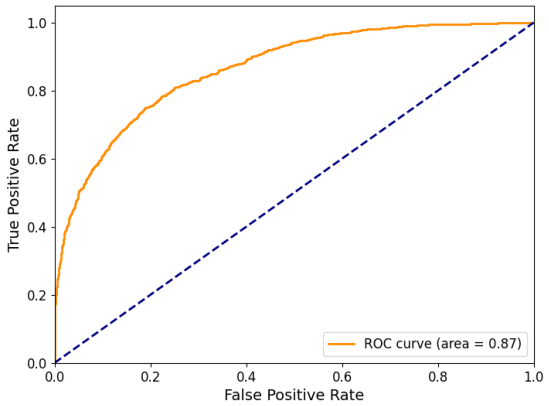
Table 4.6: Performance Metrics for Deep Learning model (ANN) for different scenarios - Adult Income dataset

Metric/Scenario	**Train on Real, Test on Real**	**Train on Synthetic, Test on Synthetic**	**Train on Synthetic, Test on Real**	**Train on Synthetic_balanced, Test on Real**
Accuracy (%)	84.96	81.12	75.84	81.95
Precision (%)	70.47	71.76	49.85	59.40
Recall (%)	64.54	62.25	61.86	78.95
F1 Score (%)	67.38	66.67	55.21	67.80
AUC_ROC (%)	90.17	86.53	79.01	90.04
Specificity (%)	91.43	89.33	80.27	82.90

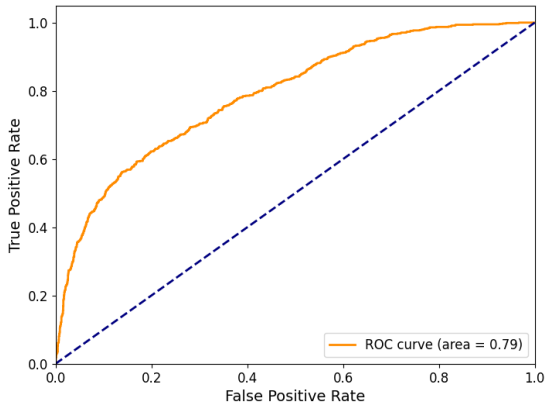
As shown in Figure 4.17, we see the change in ROC curves for the different scenarios. We see that with entire synthetic data, the AUC-ROC reduces a bit, and when training with synthetic, and testing with real, it reduces further. However, for balanced synthetic data training, the AUC-ROC is restored to that with entirely real data.



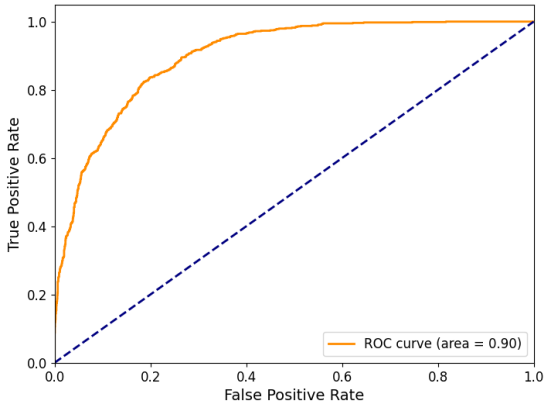
(a) Train on Real, test on Real



(b) Train on Synthetic, test on Synthetic



(c) Train on Synthetic, test on Real



(d) Train on balanced Synthetic, test on Real

Figure 4.17: ROC curves for different training and testing scenarios - Adult Income Dataset

4.2.2 MIMIC Dataset

The Machine Learning Efficacy of synthetic data for the MIMIC dataset is shown in Table 4.7. The comparative analysis of various machine learning models trained on synthetic data

and subsequently tested on real data reveals significant insights into the models' generalization capabilities and their sensitivity to class distribution changes. Our findings indicate that the generalization capability from synthetic to real data varies significantly across different models. Notably, XGBoost exhibited the largest decline in performance metrics, particularly in recall and specificity. This substantial decrease suggests a sensitivity to the over-representation of Class 0 and under-representation of Class 1 in the synthetic data. Conversely, Logistic Regression and SVM demonstrated remarkable resilience, with slight improvements in several performance metrics. These improvements suggest that these models are less affected by the slight shifts in class distribution, showcasing a robustness that makes them suitable for applications where class distribution between synthetic and real data may not align perfectly. The results underscore the critical role of model selection and synthetic data preparation in achieving effective generalization to real-world data. Models with inherent robustness to class distribution changes (e.g., Logistic Regression, SVM) are preferred in scenarios where exact class ratio replication is challenging.

Table 4.7: MLE for binary classification models for MIMIC Dataset

Model	Accuracy Difference (%)	Precision Difference (%)	Recall Difference (%)	F1 Score Difference (%)	Specificity Difference (%)
XGBoost	-0.1574 (-16.9%)	-0.0485 (-5.1%)	-0.2325 (-26.5%)	-0.16 (-17.5%)	-0.2249 (-24.5%)
Logistic Regression	0.0019 (0.24%)	0.0335 (4.92%)	-0.0121 (-1.64%)	0.0117 (1.65%)	-0.0177 (-2.09%)
Decision Tree	-0.0037 (-0.41%)	-0.0331 (-3.35%)	0.0089 (1.14%)	-0.0078 (-0.89%)	0.0146 (1.77%)
Random Forest	-0.0082 (-0.91%)	-0.047 (-5.07%)	0.013 (1.56%)	-0.0149 (-1.69%)	0.0159 (1.79%)
MLP	-0.0008 (-0.08%)	-0.0006 (-0.06%)	-0.0012 (-0.14%)	-0.001 (-0.11%)	-0.0009 (-0.10%)
SVM	0.0041 (0.45%)	-0.002 (-0.21%)	0.0092 (1.09%)	0.0043 (0.48%)	0.0079 (0.89%)

The performance metrics of deep learning model for different training and testing scenario for MIMIC dataset are listed in Table 4.8.In all real data scenario, the model exhibits excellent performance across all metrics when both trained and tested on real data. With an accuracy of 92.40%, precision at 86.12%, and an impressive recall of 95.57%, the model demonstrates its robustness in handling real-world data. The F1 Score of 90.60% indicates a strong balance between precision and recall, while the AUC-ROC of 97.62% signifies superior ability in distinguishing between classes. The specificity metric at 90.44% further confirms the model's effectiveness in correctly identifying negative cases.

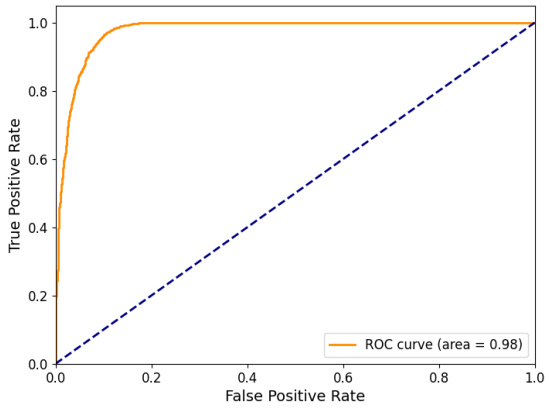
When trained and tested on synthetic data, the model's performance declines notably from the real-data scenario. Accuracy reduces to 78.70% and precision to 69.57%, reflecting challenges in accurate predictions. A significant drop in recall to 53.02% shows the model's struggles with identifying true positives. The F1 Score and AUC-ROC decrease

to 60.18% and 84.21%, indicating overall diminished effectiveness. Yet, with a specificity of 89.90%, the model retains its ability to correctly identify true negatives. the scenario where model is trained on Synthetic Data and tested on real data, tests the model's ability to generalize from synthetic to real data. Here, accuracy slightly decreases to 78.50%, but precision sees a significant increase to 89.36%, highlighting an improved positive predictive value.
The recall, however, drops to 49.80%, indicating the model struggles to identify true positives in real data. The F1 Score and AUC-ROC at 63.96% and 89.54% suggest moderate overall performance. Notably, specificity increases to 96.32%, showing excellent performance in identifying true negatives.In the last scenario, balancing the synthetic data results in improved model performance when tested on real data. Accuracy increases to 89.95%, and precision to 80.34%, indicating better overall predictive performance. Remarkably, recall jumps to 97.65%, showing the model's enhanced ability to identify true positives. The F1 Score and AUC-ROC also see increases to 88.15% and 96.87%, reflecting improved balance and discriminatory power. However, specificity drops to 85.16%, suggesting a slight decrease in the model's ability to identify true negatives compared to the real-data scenario.

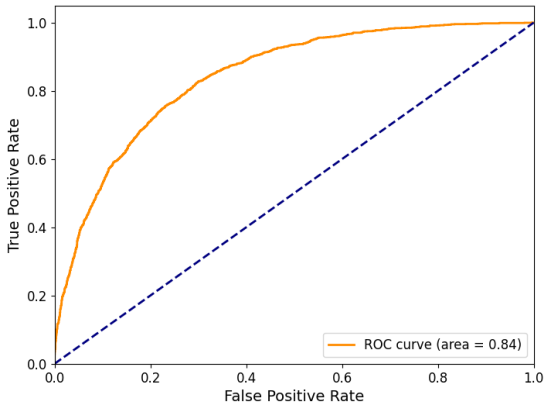
Table 4.8: Performance Metrics for Deep Learning models for different scenarios - MIMIC dataset

Metric/Scenario	**Train on Real, Test on Real**	**Train on Synthetic, Test on Synthetic**	**Train on Synthetic, Test on Real**	**Train on Synthetic_balanced, Test on Real**
Accuracy (%)	92.40	78.70	78.50	89.95
Precision (%)	86.12	69.57	89.36	80.34
Recall (%)	95.57	53.02	49.80	97.65
F1 Score (%)	90.60	60.18	63.96	88.15
AUC_ROC (%)	97.62	84.21	89.54	96.87
Specificity (%)	90.44	89.90	96.32	85.16

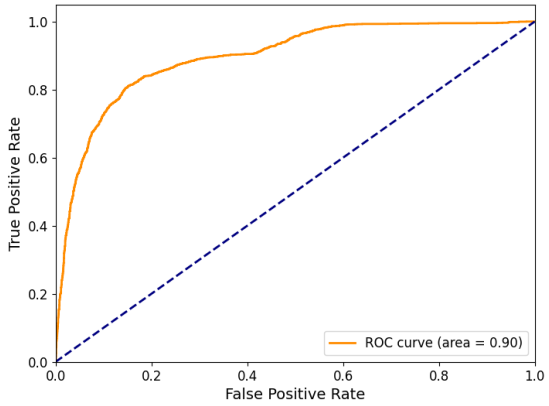
As shown in Figure 4.18, we see the change in ROC curves for the different scenarios. We see that with entire synthetic data, the AUC-ROC reduces a bit, and when training with synthetic, testing with real, it increases. However, for training balanced synehtic data and testing with real data, the AUC-ROC is restored to almost that with entirely real data.



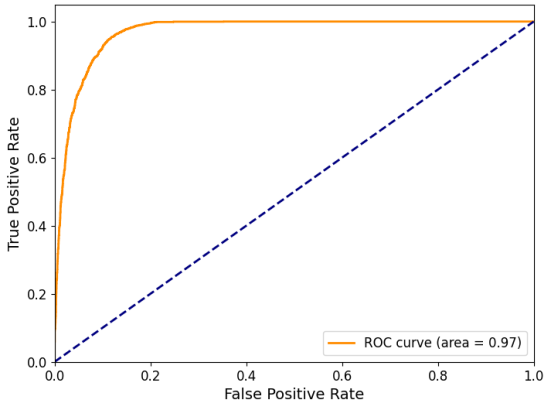
(a) Train on Real, test on Real



(b) Train on Synthetic, test on Synthetic



(c) Train on Synthetic, test on Real



(d) Train on balanced Synthetic, test on Real

Figure 4.18: ROC curves for different training and testing scenarios - MIMIC Dataset

4.2.3 Diabetes Dataset

Table 4.9 shows the Machine Learning Efficacy of Diabetes dataset for different ML models. The comparative performance analysis of machine learning models trained on synthetic versus real Diabetes Dataset reveals nuanced differences directly linked to minor variances in class distribution. The Decision Tree model shows negligible performance changes, demonstrating robustness to class ratio shifts, while Logistic Regression exhibits significant variability, notably a 15.51% increase in Precision, likely due to its sensitivity to the slight increase in Class 1 representation in the synthetic data. Other models like XG-

Boost, Random Forest, and SVM display moderate discrepancies, hinting at their varied response to class distribution adjustments. This underscores the synthetic data's potential utility, contingent on careful model selection and tuning to account for its impact on performance metrics, emphasizing the importance of aligning synthetic data closely with real data distributions for effective training outcomes.

Table 4.9: Machine Learning Efficacy (MLE) for binary classification models for Diabetes Dataset

Model	Accuracy Difference (%)	Precision Difference (%)	Recall Difference (%)	F1 Score Difference (%)	Specificity Difference (%)
XGBoost	-0.016 (-1.75%)	0.0036 (0.37%)	-0.0273 (-3.10%)	-0.0135 (-1.47%)	-0.0357 (-4.11%)
Logistic Regression	-0.0255 (-3.73%)	0.107 (15.51%)	-0.0574 (-8.43%)	0.0142 (2.07%)	-0.1578 (-23.32%)
Decision Tree	0 (0%)	0.0009 (0.09%)	-0.0005 (-0.06%)	1E-04 (0.01%)	-0.0008 (-0.10%)
Random Forest	-0.0065 (-0.72%)	-0.002 (-0.21%)	-0.0084 (-0.99%)	-0.0056 (-0.62%)	-0.0111 (-1.35%)
SVM	-0.0083 (-0.93%)	0.0023 (0.25%)	-0.0144 (-1.67%)	-0.0068 (-0.76%)	-0.0189 (-2.22%)

The performance indicators of a deep learning model for different training and testing settings on the Diabetes dataset are outlined in Table 4.10.In the all real data scenario, the model achieves an accuracy of 87.87% when trained and tested on real data, indicating a relatively high overall performance. However, the precision (25.98%) and recall (4.67%) are low, highlighting a challenge in accurately identifying the minority class (Class 1) in the presence of a substantial majority class (Class 0), as shown by the very high specificity (98.33%). This suggests that the model is biased towards predicting the majority class correctly at the expense of the minority class, which is also reflected in a modest F1 Score (7.91%) and AUC-ROC (62.44%).

For the all synthetic data scenario, accuracy slightly improves to 88.44%, indicating that the model performs well in the synthetic domain where the class distribution closely mirrors that of the real data. However, precision, recall, and F1 Score all drop to 0.00%, showing the model's complete failure to identify the minority class within the synthetic environment. This is further evidenced by a specificity of 100%, demonstrating a perfect identification of the majority class while completely missing the minority class, leading to a decreased AUC-ROC (59.32%).When trained on synthetic data and tested on real data, the model maintains a relatively high accuracy (86.73%). Yet, the precision (12.06%), recall (2.99%), and F1 Score (4.80%) are extremely low, underscoring significant difficulties in generalizing the detection of the minority class from synthetic to real settings. This scenario highlights the model's inclination towards the majority class, as indicated by a high specificity (97.26%) but poor ability to distinguish the minority class, as shown by a lower AUC-ROC (52.90%).

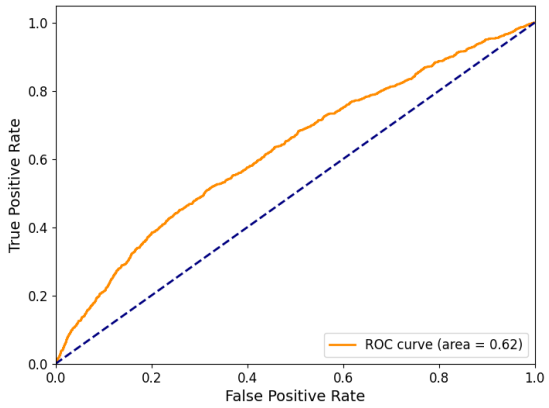
Finally, adjusting the synthetic data to balance the class ratio leads to notable improvements. The accuracy stands at 87.04%, with significant enhancements in precision (41.58%), recall (39.79%), and F1 Score (40.67%), indicating a marked improvement in identifying the minority class. The rise in AUC-ROC to 78.72% reflects better discrimination between classes. Specificity decreases to 92.98%, suggesting a reduced bias towards the majority class and a more balanced approach in predicting both classes. This scenario demonstrates the critical impact of class ratio adjustments in synthetic data on model performance, especially for minority class detection.

These modifications underscore the influence of class distribution in training datasets on the predictive performance of models, particularly in relation to class imbalance. Adjusting the synthetic data to more closely reflect or balance the class distribution proves crucial for enhancing model sensitivity towards the minority class and achieving a more balanced prediction capability.

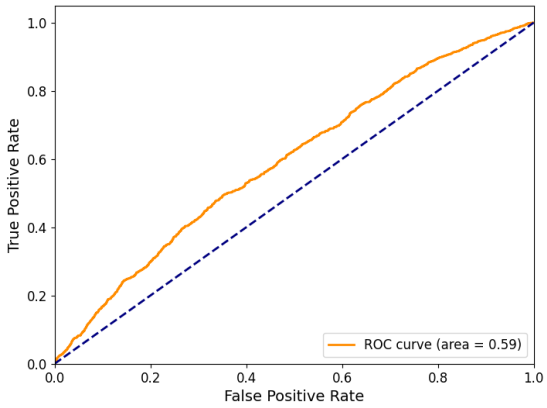
Table 4.10: Performance Metrics for Deep Learning models for different scenarios - Diabetes Dataset

Metric/Scenario	**Train on Real, Test on Real**	**Train on Synthetic, Test on Synthetic**	**Train on Synthetic, Test on Real**	**Train on Synthetic_balanced, Test on Real**
Accuracy (%)	87.87	88.44	86.73	87.04
Precision (%)	25.98	0.00	12.06	41.58
Recall (%)	4.67	0.00	2.99	39.79
F1 Score (%)	7.91	0.00	4.80	40.67
AUC_ROC (%)	62.44	59.32	52.90	78.72
Specificity (%)	98.33	100.00	97.26	92.98

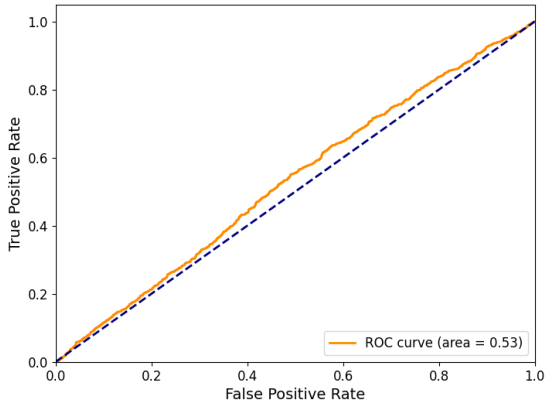
As shown in Figure 4.19, we see the change in ROC curves for the different scenarios for Diabetes dataset. We see that with entire synthetic data, the AUC-ROC reduces by (around 3% value), and when training with synthetic, testing with real, it reduces even further. However, for training balanced synthetic data and testing with real data, the AUC-ROC enhances and even surpasses to that with entire data. This exhibits the use of synthetic data for data augmentation and improving the performance of model by balancing the classes.



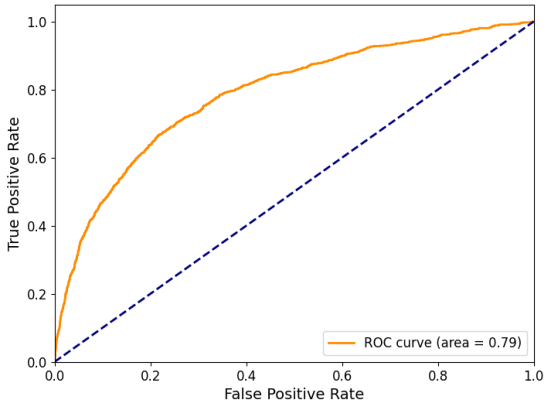
(a) Train on Real, test on Real



(b) Train on Synthetic, test on Synthetic



(c) Train on Synthetic, test on Real



(d) Train on balanced Synthetic, test on Real

Figure 4.19: ROC curves for different training and testing scenarios - Diabetes Dataset

4.2.4 Synthetic Data for Augmentation of Real Data

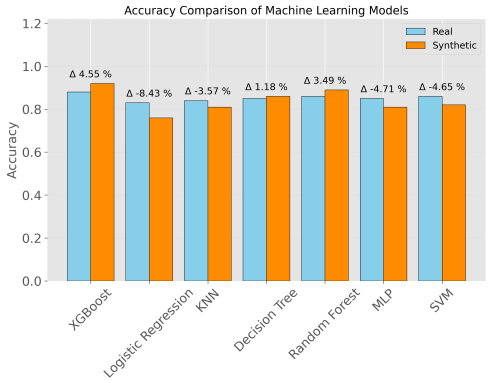
Synthetic Data has also been commonly used to augment the data to enhance the performance of the predictive models [31]. We conducted experiments to study the change in the metrics for different combinations of synthetic and real data.

Since class ratio of a dataset plays a crucial role in the performance of a machine learning model, a study is done to compare the performance for the real data with original class ratio and the real data augmented with synthetic data that makes the class ratio 1:1. The performances of the ML models are compared in Table 4.11 and shown in Figure 4.20.

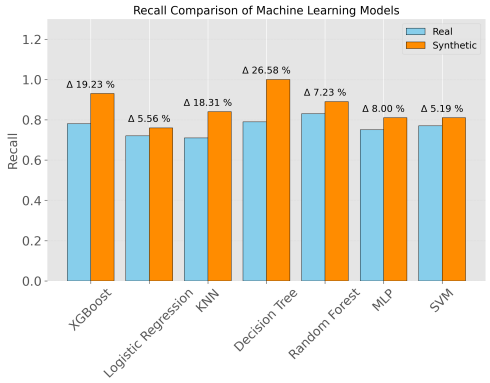
The accuracy changes were minimal across most models, with XGBoost showing a slight improvement of 4.55% and Logistic Regression experiencing the most significant decrease of -8.43%. This variance indicates that while balancing class distribution through synthetic data augmentation can enhance model accuracy in certain cases (e.g., XGBoost), it might detriment others (e.g., Logistic Regression). Notably, all models saw an improvement in precision, with Random Forest exhibiting the highest increase of 62.96%. This substantial precision enhancement suggests that synthetic data augmentation effectively addresses the class imbalance issue, leading to a higher rate of correct positive predictions across models.Recall scores varied, with the Decision Tree model benefiting the most, showing a 26.58% improvement. F1 scores, which harmonize the precision and recall metrics, generally saw improvements, with the Decision Tree model again standing out with a 33.33% increase. These improvements suggest that for certain models, synthetic data augmentation can enhance the overall balance between precision and recall. Specificity saw declines in several models, with Logistic Regression marking the most notable decrease of -19.15%.

Table 4.11: Performance difference for Augmentation using Synthetic Data (Class Ratio = 1:1) for Adult Dataset

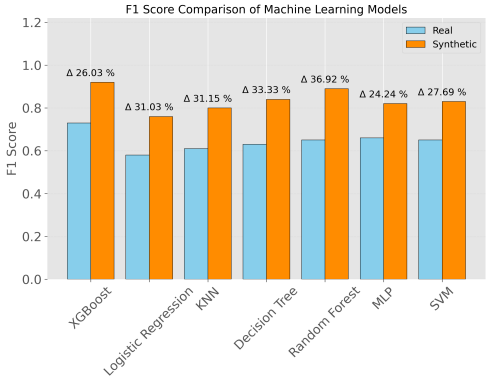
Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**Specificity**
XGBoost	0.04 (4.55%)	0.21 (30.43%)	0.15 (19.23%)	0.19 (26.03%)	-0.01 (-1.06%)
Logistic Regression	-0.07 (-8.43%)	0.26 (53.06%)	0.04 (5.56%)	0.18 (31.03%)	-0.18 (-19.15%)
KNN	-0.03 (-3.57%)	0.25 (47.17%)	0.13 (18.31%)	0.19 (31.15%)	-0.08 (-8.6%)
Decision Tree	0.01 (1.18%)	0.21 (40.38%)	0.21 (26.58%)	0.21 (33.33%)	0.04 (4.17%)
Random Forest	0.03 (3.49%)	0.34 (62.96%)	0.06 (7.23%)	0.24 (36.92%)	-0.07 (-7.22%)
MLP	-0.04 (-4.71%)	0.23 (38.98%)	0.06 (8.0%)	0.16 (24.24%)	-0.14 (-14.89%)
SVM	-0.04 (-4.65%)	0.29 (51.79%)	0.04 (5.19%)	0.18 (27.69%)	-0.15 (-15.79%)



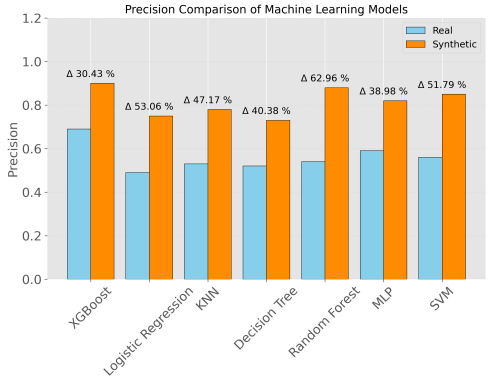
(a) Accuracy



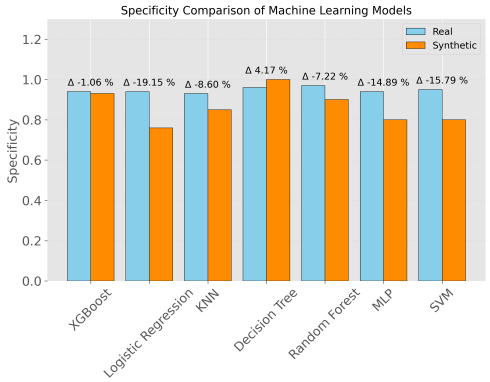
(b) Recall



(c) F1 Score



(d) Precision



(e) Specificity

Figure 4.20: Change in Metrics for Augmentation using Synthetic Data (Class Ratio = 1:1) for Adult Dataset

Balancing classes through synthetic data augmentation has shown to notably improve precision across all machine learning models tested, with effects on other metrics being model-dependent. This strategy contributes to creating more generalized models by mitigating the skewness inherent in imbalanced datasets. Despite mixed impacts on accuracy, recall, and specificity, the overall trend towards improved model robustness and generalizability highlights the value of class balancing in enhancing machine learning performance.

4.3 Privacy Metrics - Reidentification Risks

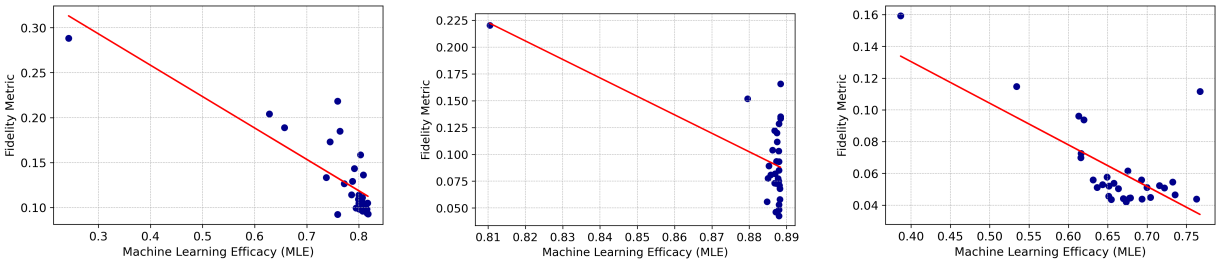
The outcomes of the privacy analysis are indicative of the efficacy of the adopted de-identification procedures across the three distinct datasets under scrutiny. The evaluation employed the reidentification risk assessment model proposed by Khaled et al., which facilitates a nuanced appraisal of the privacy risks associated with both real and synthetic tabular data. The calculated reidentification scores serve as a testament to the robustness of the privacy-preserving methodologies implemented. For the Adult Income dataset, a score of 0.00024 was obtained, whereas the MIMIC dataset was associated with a reidentification score of 0.0326. The Diabetes dataset demonstrated a score of 0.00098. It is noteworthy that all these values fall below the acceptable risk threshold set forth by the European Medicines Agency and Health Canada, denoted as 0.09 in this context. The results thus underscore a satisfactory alignment with regulatory standards for privacy, substantiating the datasets' readiness for subsequent use in research while ensuring the maintenance of individual privacy.

4.4 Hyperparamater Tuning for the Synthetic Data Generation

For developing a use-case agnostic framework of HPT of synthetic data generation, the correlation between Utility and Fidelity metrics for different hyperparameter combinations is studied and shown in Table 4.12.

Table 4.12: Correlation between Fidelity and Utility Metrics for HPT of SDG

Datasets					
Adult Income		**Diabetes**		**MIMIC**	
Pearson	**p-value**	**Pearson**	**p-value**	**Pearson**	**p-value**
-0.81	5.01E-08	-0.61	3.00E-04	-0.69	2.35E-05



(a) Adult Income (b) Diabetes (c) MIMIC

Figure 4.21: Correlation of Fidelity and Utility Metrics

As shown in Figure 4.21 and listed in Table 4.12, there is a notable negative corrrelation between the fidelity and utility metrics for the hyperparameter tuning of the synthetic data generation process across all datasets, indicating that as one metric increases, the other tends to decrease. For fidelity metrics (HD and DPC), lower values are desirable, which represent the statistical closeness of the real and synthetic data.

On the other hand, larger values of MLE metrics are desirable, since they represent the performance of the synthetic data to train machine learning models in replacement of real data. The Adult Income dataset exhibits negative correlation (Pearson coefficient = -0.81) with a significant p-value (5.01E-08), suggesting a robust inverse relationship between the two metrics. The Diabetes dataset shows a moderate negative correlation (Pearson coefficient = -0.61) with a significant p-value (3.00E-04), indicating a notable but less pronounced inverse relationship compared to the Adult Income dataset. The MIMIC dataset presents a negative correlation (Pearson coefficient = -0.69) with a significant p-value (2.35E-05), which is stronger than the Diabetes dataset but still less than the Adult Income dataset. The significant p-values across all datasets represent the statistical significance of the observed correlations.

4.5 Hyperparameter Tuning of Predictive Models using Synthetic Data

4.5.1 Mortality Prediction- HPT Binary Classification (ANN) - MIMIC Dataset

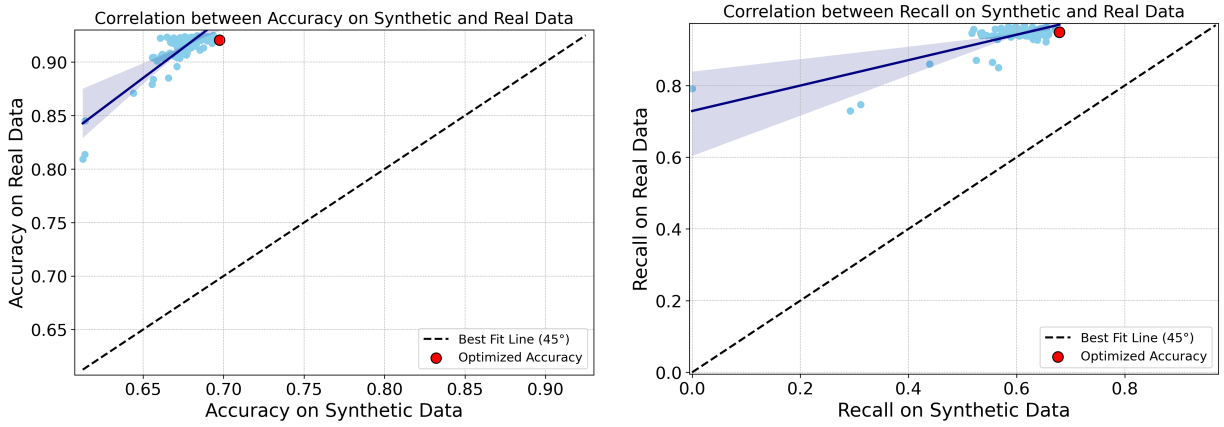
Figure 4.13 shows the comparison of Pearson correlation coefficients and p-values for the MIMIC dataset when tuning the hyperparameters using the synthetic data and then using the same parameters to evaluate the performance on the real data. To confirm the hypothesis, the experiments are also applied to synthetic data generated using a transformer-based model - RealTabFormer.

Table 4.13: Pearson correlation coefficients and p-values for MIMIC dataset HPT using Synthetic data for a feed-forward ANN binary classifier

MIMIC dataset - Binary Classification - ANN				
	CTGAN		**RealTabFormer**	
Metric	**Pearson**	**p-value**	**Pearson**	**p-value**
Accuracy	0.86	1.07E-30	0.89	4.42E-18
Precision	0.41	1.86E-05	0.86	2.05E-15
Recall	0.77	3.78E-21	0.53	8.52E-05
F1	0.75	3.71E-19	0.87	2.35E-16

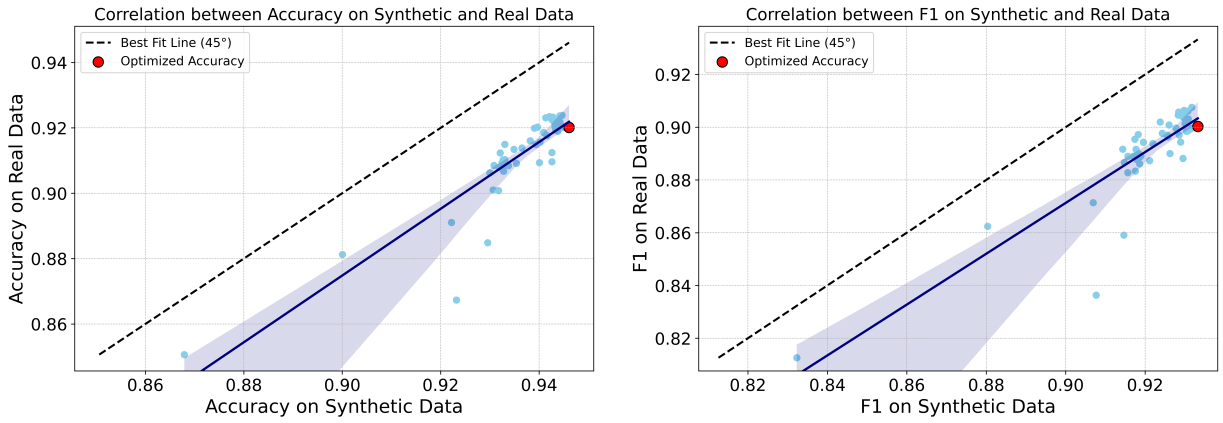
For both synthetic data generation methods, most metrics show a high Pearson correlation coefficient with real data performance, particularly in the cases of Accuracy and F1 Score. This suggests a strong linear relationship between the performance metrics achieved with synthetic data and those achieved with real data, indicating that synthetic data can be effectively used for hyperparameter tuning in deep learning models. Both methods demonstrate high Pearson correlation coefficients for Accuracy and F1 Score, with RealTabFormer slightly outperforming CTGAN. The Pearson correlation coefficients and p-values for performance metrics using synthetic data in the MIMIC dataset reveal insightful trends for hyperparameter tuning in a feed-forward ANN binary classifier.
Accuracy demonstrates a high correlation (0.86 for CTGAN and 0.89 for RealTabFormer), indicating synthetic data's effectiveness in mirroring overall classification performance. Precision varies markedly between CTGAN (0.41) and RealTabFormer (0.86), highlighting the dependency of synthetic data quality on the generation method, with RealTabFormer

being particularly adept at replicating the positive predictive value. Recall shows a moderate to high correlation (0.77 for CTGAN and 0.53 for RealTabFormer), pointing out the challenges in accurately capturing the model's sensitivity using synthetic data. The F1 Score, balancing precision and recall, also exhibits high correlation (0.75 for CTGAN and 0.87 for RealTabFormer), underscoring synthetic data's capability to simulate the real data's balanced metric performance. These findings collectively underscore the potential of synthetic data, to effectively replicate real dataset characteristics across various performance metrics, making it a valuable tool for model development and optimization.



(a) Accuracy (b) Recall

Figure 4.22: Correlation of performance on synthetic and real data - CTGAN - MIMIC



(a) Accuracy

(b) F1 score

Figure 4.23: Correlation of performance on synthetic and real data - RealTabFormer - MIMIC

Figure 4.22 and 4.23 show the correlation graphs for the metrics with the best correlations between real and synthetic data hyperparameter tuning. The 45 degrees line from the origin shows the best correlation. The inferences from these graphs are two-fold, first, we can study the correlation between the hyperparameter tuning using synthetic and real data, and second, we can see the change in the metrics when a model is trained on real data compared to when it is trained on synthetic data (Machine Learning Efficacy). Additionally, the red point represents the metric value for the best hyperparameter combinations when hyperparameter tuning is done on synthetic data. HPT aims to find the best hyperparameters that give the best performance metrics. Therefore, this point is a good measure of the correlation study between real and synthetic data for HPT. The optimized metric for HPT using synthetic data should correlate with the optimized metric for HPT using real data.

In Figure 4.22 (a), the correlation line is parallel to the best-fit line which indicates a strong correlation, which is also evident by the Pearson coefficient of 0.86 and p-value of 1.07e-3. Secondly, since the correlation line is a bit far from the best-fit line, it indicates the reduction of the accuracy when the model is trained on synthetic data compared to when trained on real data. In Figure 4.22 (b), the observed correlation is of a moderate level, indicating a balanced relationship. The observed reduction in correlation is not considered to be of significant concern for the analysis. In both of the figures, the optimized metric for the synthetic data HPT is located at the top of the correlation line and also corresponds

to the largest metric for the real data HPT.

4.5.2 Readmission Prediction- HPT Binary Classification (ANN) - Diabetes Dataset

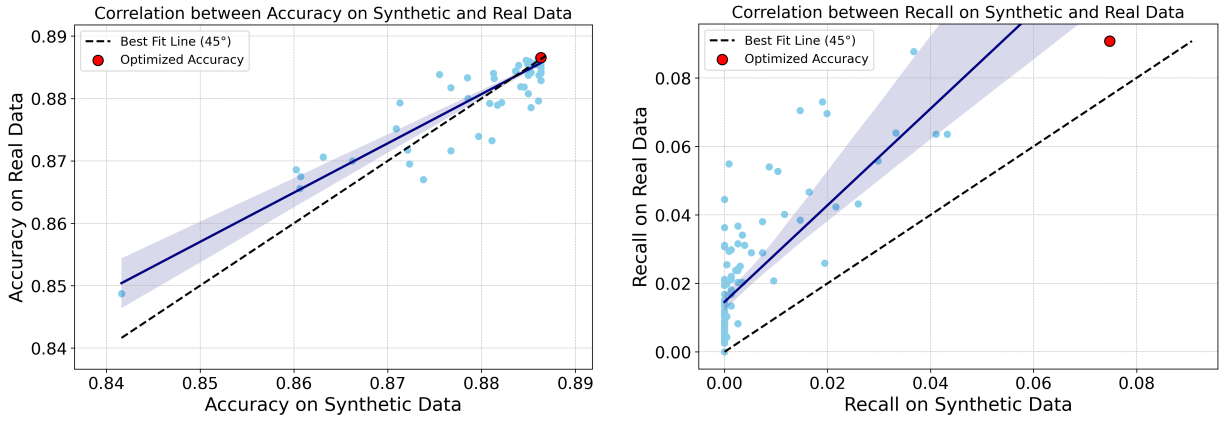
4.14 shows the comparison of Pearson correlation coefficients and p-values for the Diabetes dataset when tuning the hyperparameters using the synthetic data and then using the same parameters to evaluate the performance on the real data.

Table 4.14: Pearson correlation coefficients and p-values for Diabetes dataset HPT using Synthetic data for a feed-forward ANN binary classifier

Diabetes - Binary Classification - ANN				
	CTGAN		**RealTabFormer**	
Metric	**Pearson**	**p-value**	**Pearson**	**p-value**
Accuracy	0.92315	1.78E-42	0.94667	2.96E-25
Precision	-0.204	0.041771	-0.2217	1.22E-01
Recall	0.79128	1.17E-22	0.85285	3.80E-15
F1	0.76588	1.67E-20	0.80498	1.85E-12

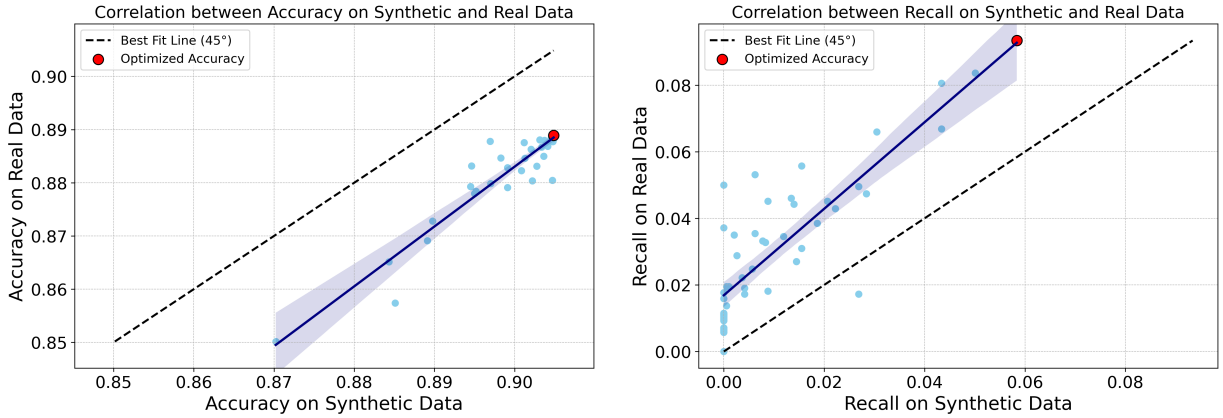
Both CTGAN and RealTabFormer demonstrate strong positive Pearson correlation coefficients with respect to accuracy (0.92315 and 0.94667, respectively), suggesting that hyperparameters optimized on synthetic data translate well to real data, leading to high accuracy in the ANN binary classifier. The extremely low p-values indicate a statistically significant correlation, reinforcing the effectiveness of using synthetic data (generated by both CTGAN and RealTabFormer) for HPT.
The negative correlation coefficients for precision (-0.204 with CTGAN and -0.2217 with RealTabFormer) might initially seem counterintuitive. However, given the hypothesis, this suggests that there may be a disparity in how well the precision-focused hyperparameters optimized on synthetic data perform when applied to real data. Positive correlation coefficients for recall (0.79128 with CTGAN and 0.85285 with RealTabFormer) indicate that hyperparameters optimized for recall on synthetic data are effective when applied to real data. Similarly, the strong positive Pearson correlation coefficients for the F1 score (0.76588 with CTGAN and 0.80498 with RealTabFormer) suggest that hyperparameters that optimize the F1 score on synthetic data have a similar optimizing effect on real data.

Figures 4.24 and 4.25 show the visual representation of the metrics with the best correlations for the Diabetes dataset.



(a) Accuracy (b) Recall

Figure 4.24: Correlation of performance on synthetic and real data - CTGAN - Diabetes



(a) Accuracy (b) Recall

Figure 4.25: Correlation of performance on synthetic and real data - RealTabFormer - Diabetes

4.5.3 Income Category Prediction - HPT Binary Classification (ANN) - Adult Income Dataset

4.15 shows the comparison of Pearson correlation coefficients and p-values for the Diabetes dataset when tuning the hyperparameters using the synthetic data and then using the

same parameters to evaluate the performance on the real data.

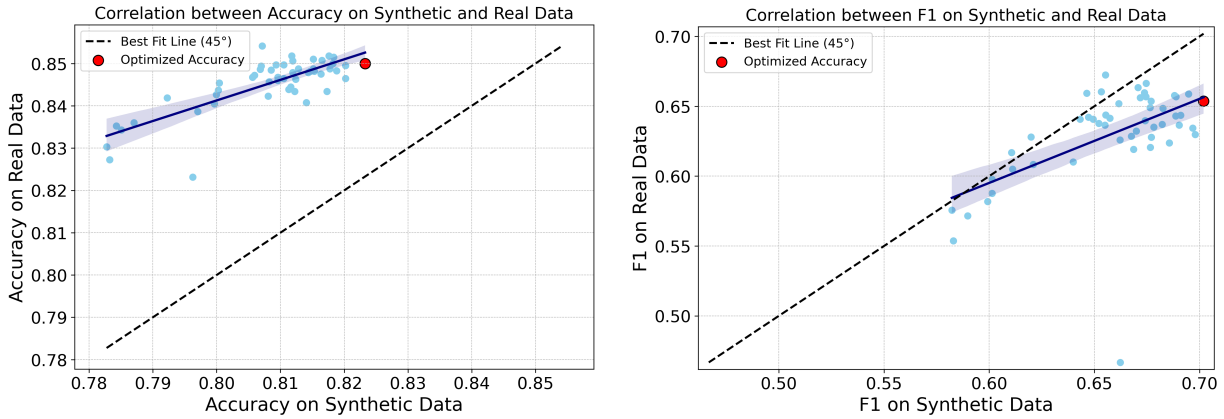
Table 4.15: Pearson correlation coefficients and p-values for Adult Income dataset HPT using Synthetic data for a feed-forward ANN binary classifier

Adult Income - Binary Classification - ANN				
	CTGAN		**RealTabFormer**	
Metric	**Pearson**	**p-value**	**Pearson**	**p-value**
Accuracy	0.79	1.09E-11	0.96	7.79E-31
Precision	-0.28	0.053	0.97	9.85E-32
Recall	0.35	0.013	0.97	3.48E-35
F1	0.55	3.08E-05	0.97	2.91E-33

The Pearson correlation coefficients and p-values for performance metrics using synthetic data in the Adult dataset, as applied in hyperparameter tuning for a feed-forward ANN binary classifier, present a nuanced picture of the utility and effectiveness of synthetic data. Accuracy displays strong correlations for both methods, particularly high for RealTabFormer (0.96), suggesting an excellent capability of synthetic data to predict overall classification success. The significant correlation, coupled with extremely low p-values, confirms the statistical robustness of these results, indicating that synthetic data, especially from RealTabFormer, can closely mimic real-world data outcomes in terms of overall model accuracy.

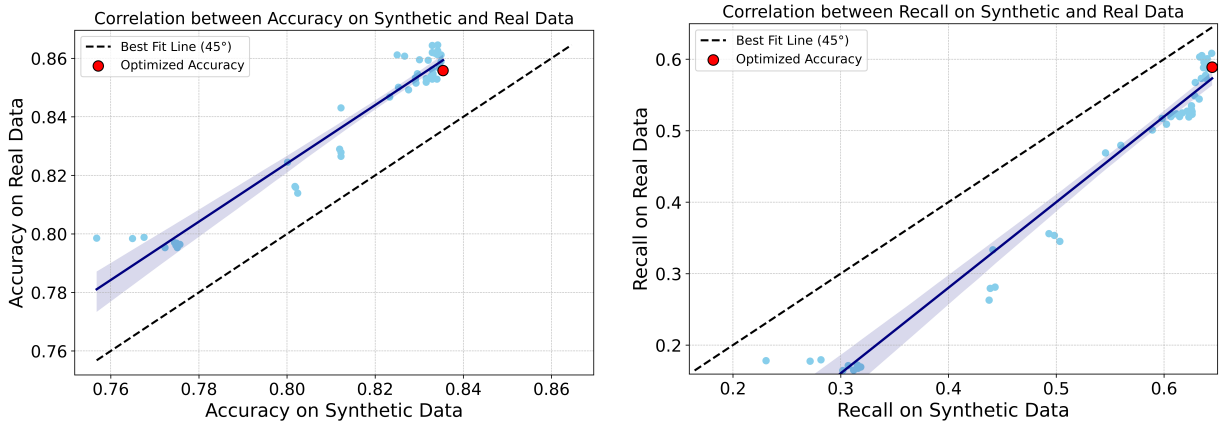
Precision and recall metrics underscore the disparity in performance between CTGAN and RealTabFormer with synthetic data. CTGAN's precision shows a slight negative correlation (-0.28), hinting at challenges in mimicking real data's positive predictive value, whereas RealTabFormer excels with a near-perfect correlation (0.97), effectively reflecting real data precision in synthetic datasets. On the recall front, CTGAN achieves a moderate positive correlation (0.35), showing some effectiveness in model sensitivity. Conversely, RealTabFormer's recall correlation soars at 0.97, indicating its superior capability in accurately identifying true positives. F1 Score, which harmonizes precision and recall, underscores the high efficacy of RealTabFormer (0.97) in creating synthetic data that reflects the balanced metric performance seen in real data, evidenced by the statistically significant p-values. CTGAN shows a moderate correlation (0.55), indicating a somewhat effective balance but not nearly as closely aligned with real data as RealTabFormer. The stronger correlations when using RealTabFormer directly relate to the better fidelity metrics as shown earlier in Table 4.4. This indicates that synthetic data with superior fidelity metrics, implying enhanced statistical representativeness, leads to stronger correlations in hyperparameter tuning outcomes between synthetic and real data, underscoring the value

of high-quality synthetic datasets. Figures 4.26 and 4.27 shows the visual representation of the metrics with the best correlations for the Adult Income dataset.



(a) Accuracy (b) F1 Score

Figure 4.26: Correlation of performance on synthetic and real data - CTGAN - Adult Income dataset



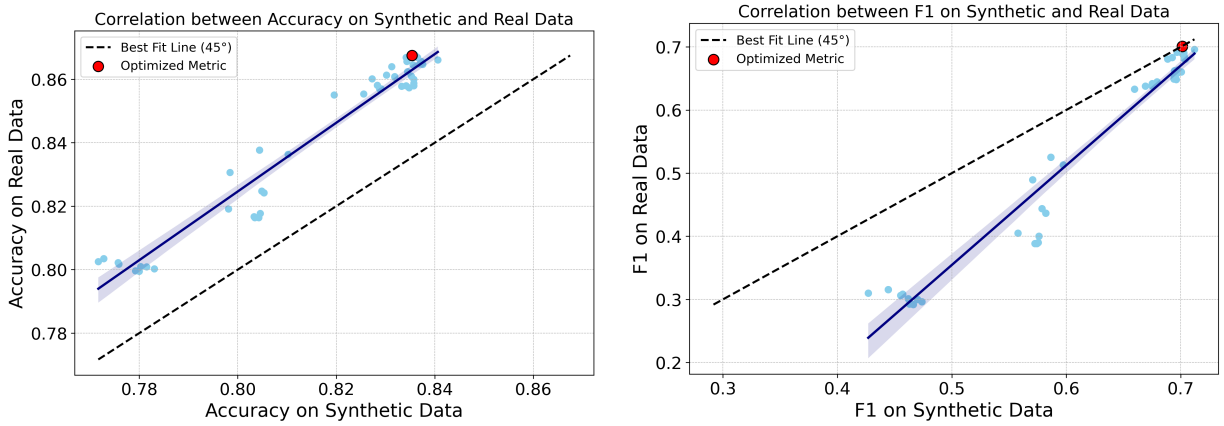
(a) Accuracy (b) Recall

Figure 4.27: Correlation of performance on synthetic and real data - RealTabFormer - Adult Income dataset

4.5.4 HPT Binary Classification (Random Forest)

Table 4.16: Adult Income - Binary Classification - Random Forest

Metric	CTGAN		RealTabFormer	
	Pearson	p-value	Pearson	p-value
Accuracy	0.97437	9.37E-33	0.99111	1.05E-43
Precision	0.93433	3.81E-23	0.96131	1.58E-28
Recall	0.97090	7.46E-35	0.99236	2.80E-45
F1	0.97645	1.25E-33	0.99294	4.14E-46



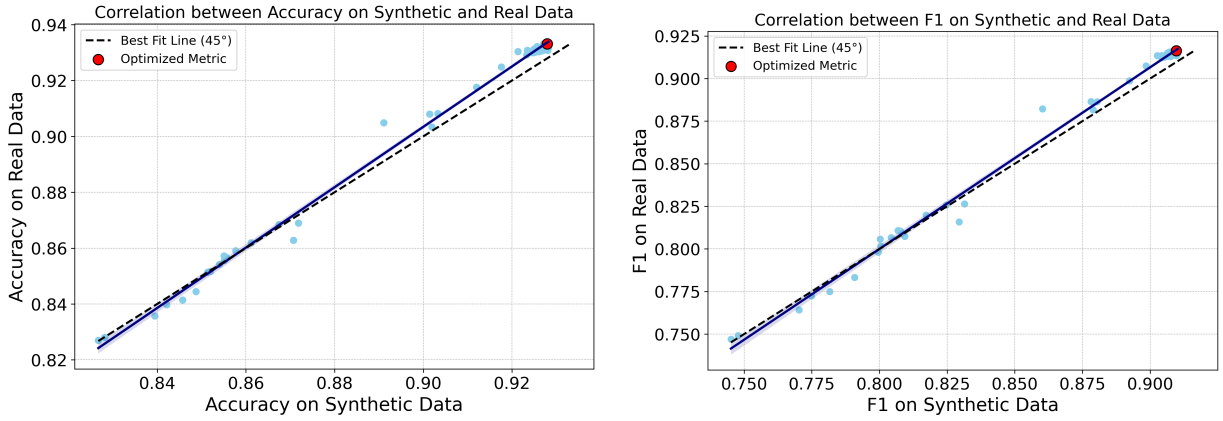
(a) Accuracy

(b) F1 Score

Figure 4.28: Correlation of performance on synthetic and real data - CTGAN - Adult Income

Table 4.17: MIMIC dataset - Binary Classification - Random Forest

Metric	CTGAN		RealTabFormer	
	Pearson	p-value	Pearson	p-value
Accuracy	0.99751	6.15E-57	0.9855	1.22E-38
Precision	0.98098	7.84E-36	0.96263	6.98E-29
Recall	0.99462	6.15E-49	0.9851	2.34E-38
F1	0.9971	2.40E-55	0.98516	2.11E-38



(a) Accuracy (b) F1 Score

Figure 4.29: Correlation of performance on synthetic and real data - CTGAN - MIMIC

4.5.5 HPT Regression model (ANN)

To validate the generalizability of the hypothesis, we perform hyperparameter tuning of a feed-forward neural network for a regression model. For this, we consider the MIMIC dataset and use the remaining variables to predict the 'length of stay' in the hospital. Figure 4.30 shows the correlation of the Root Mean Squared Error (RMSE) for the prediction. We perform another experiment using the Adult Income dataset, to predict the hours-per-week (continuous variable) using the rest variables. Figure 4.31 shows the correlation of the Root Mean Squared Error (RMSE) for the prediction.

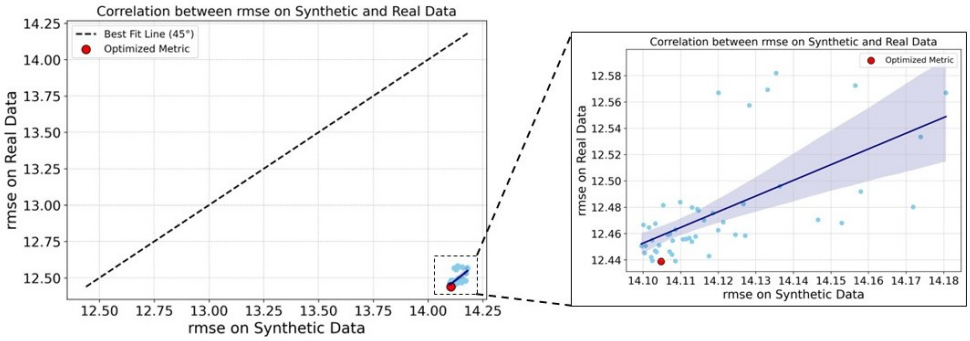


Figure 4.30: Correlation of RMSE (Root Mean Squared Error) for predicting Length of stay in hospital using MIMIC synthetic data

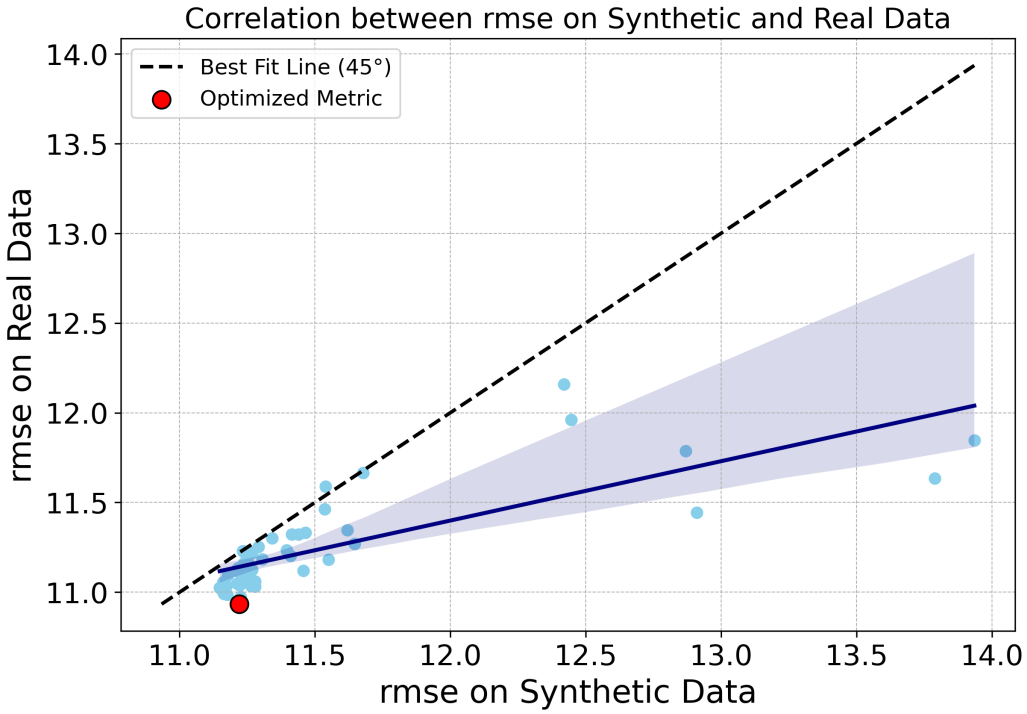


Figure 4.31: Correlation of RMSE (Root Mean Squared Error) for predicting hours-per-week using Adult Income synthetic data

In Figure 4.30, the correlation line is parallel to the best-fit line, indicating a strong correlation between the RMSE for HPT using real and synthetic data. This is also evident by a Pearson correlation coefficient of 0.63, and p-value ¡1e-8. Both datasets show strong Pearson correlation coefficients for RMSE (0.77 for Adult Income, 0.63 for MIMIC) and $R\hat{2}$ (0.75 for Adult Income, 0.76 for MIMIC), with highly significant p-values. This suggests that ANNs have a robust predictive capability in both contexts, with predictions closely aligned with actual outcomes. The relatively low RMSE values indicate good model accuracy. The MAE presents a mixed outcome, with a relatively low Pearson coefficient for Adult Income (0.34) and an even lower one for MIMIC (0.03). The p-value for Adult Income suggests that this result is statistically significant, whereas the high p-value for MIMIC indicates a lack of statistical significance. The significant p-values for RMSE and $R\hat{2}$ across both datasets reinforce the utility of synthetic data in creating effective predictive models.

Table 4.18: Continuous Prediction Efficacy Using ANN

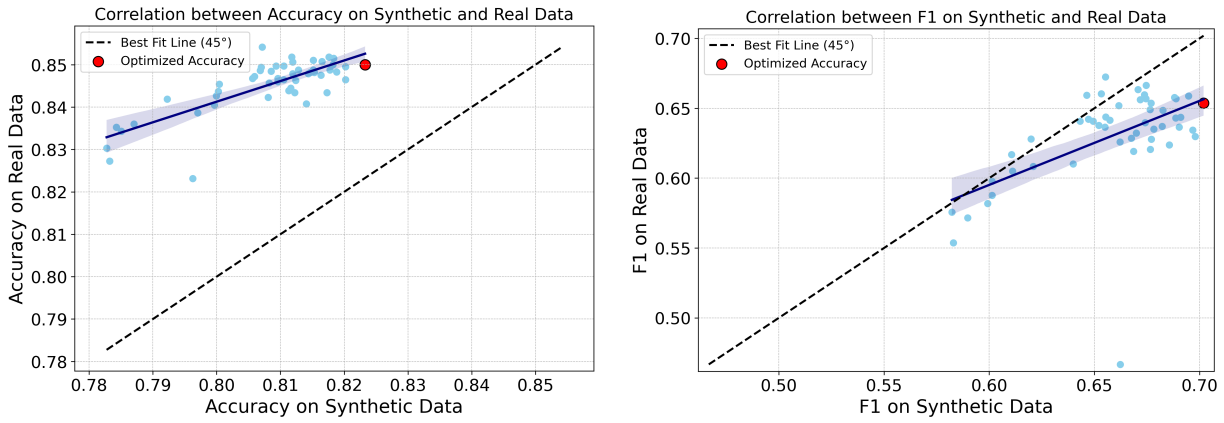
Metric	Adult Income		MIMIC	
	Pearson	**p-value**	**Pearson**	**p-value**
RMSE	0.77	5.62E-11	0.63	9.89E-07
R2	0.75	1.88E-10	0.76	1.78E-09
MAE	0.34	1.30E-02	0.03	8.18E-01

4.5.6 HPT on Real Data and Correlation with Synthetic Data (ANN)

To validate the hypothesis of applying the parameter tuning from synthetic data to real data, we reverse the process by tuning the hyperparameters on real data and then applying those parameters to the synthetic data. The results in Table 4.19 reveal that there is a strong correlation between the Accuracy, and F1 score, a moderate correlation for Recall, and a weak to moderate correlation for Precision. Figure 4.32 also shows the regression lines for the correlation of Accuracy/F1 on synthetic and real data is parallel to the best-fit line, and the best point indicating the best performance of real data corresponds to the best point when using synthetic data.

Table 4.19: Income Category prediction - Binary Classification - ANN (Real to synthetic) - Adult Income dataset

Metric	Pearson	p-value
Accuracy	0.92534	7.47E-22
Precision	0.28518	0.044702
Recall	0.62122	1.48E-06
F1	0.81035	1.01E-12



(a) Accuracy

(b) F1 Score

Figure 4.32: Correlation of performance on synthetic and real data - CTGAN - Adult Income

4.6 Effect of Stratification in Synthetic Data Generation

Since HPT is a computationally expensive process, we study the use of stratified random sampling to save computational time, and accelerate SDG. We perform the stratified random sampling approach for hyperparameter tuning of SDG to the Adult Income Dataset, based on the variable 'work-class' and reduce the dataset size from 32561 to 3000. Figure 4.33 shows the distribution of the HD for different hyperparameter combinations for original data and stratified samples. As mentioned in Table 4.20, with the stratified sample we see that we save 261 minutes of computational time for the Synthetic Data Generation hyperparametes tuning, with a difference of 0.0059 in Mean HD value. Next, to validate this hypothesis, we applied the best parameters from the stratified sample HPT to the original real data and we get a Mean HD of 0.1059. Therefore, the Mean HD reduces by a value of 0.0098, which is not a significant reduction compared to the computational time saved.

Table 4.20: Minimum Mean Hellinger Distance and Computational Time for HPT on stratified real data

Data	**Minimum Mean Hellinger Distance**	**Computational Time (Minutes)**
Original Data Size	0.0952	293
Stratified Sample	0.1011	74
Computational Time Saved		261

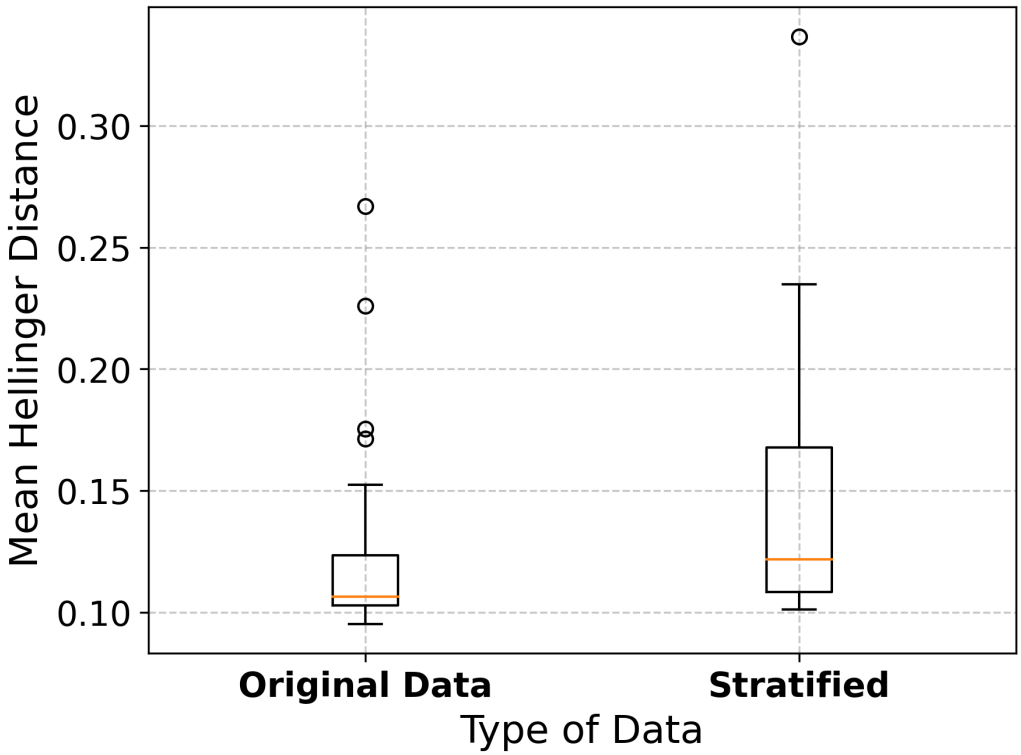


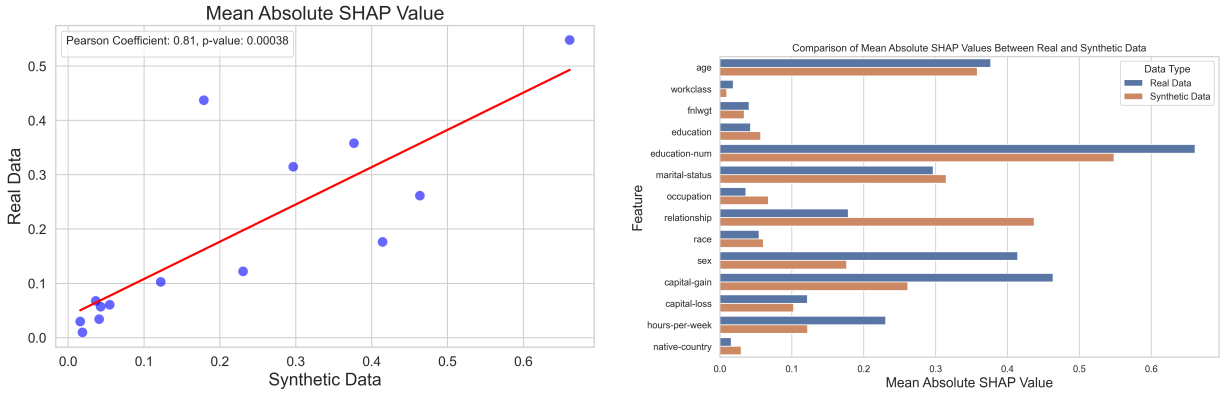
Figure 4.33: Boxplot for the Mean Hellinger Distance for HPT using Original and Stratified Data

4.7 Explainability using Synthetic Data

Another use of synthetic data is to explain the models trained on real data. Models which are trained on real data, are often shared with other organizations for testing and deployment, however, the original data is not shared. In these cases, synthetic data can help understand how these machine learning models are achieving a specific prediction. We study the correlation between Mean SHAP values of all features of synthetic and real data as shown in Table 4.21. The results show that there is a good correlation between the synthetic and real data mean SHAP values.

Table 4.21: Pearson Correlation Coefficients and p-values of Mean SHAP values between synthetic and real data

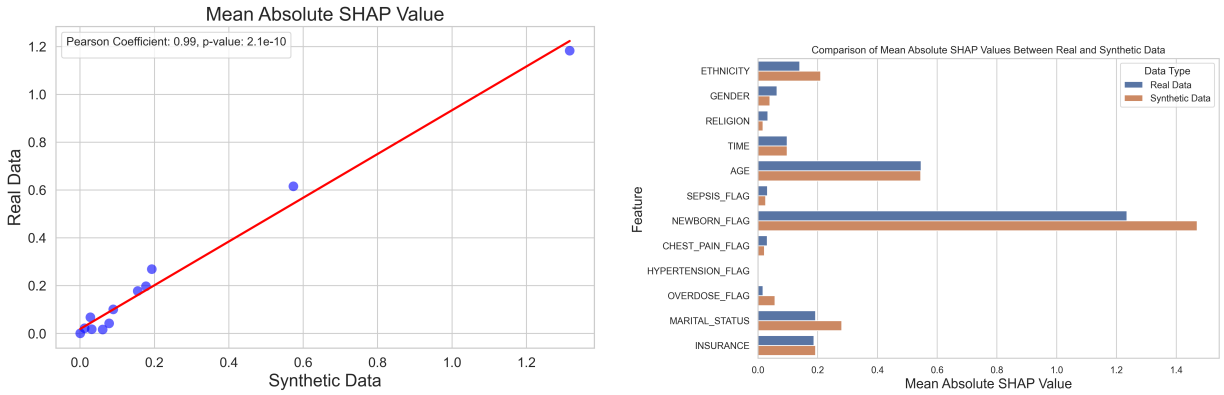
Dataset	Pearson Correlation Coefficient	P-value
MIMIC	0.99	$2.10e-10$
Adult Income	0.81	$3.84e-4$
Diabetes	0.85	$7.87e-14$



(a) Correlation

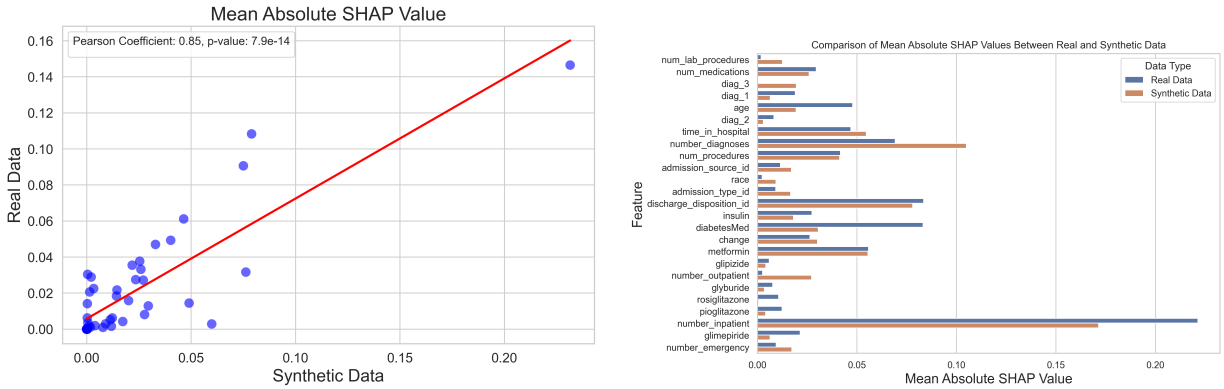
(b) Mean Absolute SHAP Values

Figure 4.34: Comparison of Mean Absolute SHAP Values Between Real and Synthetic Data - Adult Income Dataset



(a) Correlation (b) Mean Absolute SHAP Values

Figure 4.35: Comparison of Mean Absolute SHAP Values Between Real and Synthetic Data - MIMIC Dataset



(a) Correlation (b) Mean Absolute SHAP Values

Figure 4.36: Comparison of Mean Absolute SHAP Values Between Real and Synthetic Data - Diabetes Dataset

In a comprehensive analysis spanning three datasets—*MIMIC*, *Adult Income*, and *Diabetes*—Pearson correlation coefficients between mean SHAP values of features in synthetic versus real data reveal significant congruence. For *MIMIC*, a near-perfect correlation (0.99) with a p-value of $2.10e-10$ is observed. The *Adult Income* dataset demonstrates a robust correlation of 0.81 (p-value: $3.84e-4$), while the *Diabetes* dataset exhibits a very strong

correlation of 0.85 (p-value: $7.87e-14$).

These results suggest that synthetic data accurately mirrors the feature importance of real datasets across diverse fields, including medical, socioeconomic, and health research. The almost perfect correlation in the *MIMIC* dataset underscores the potential of synthetic data to serve as a reliable substitute for real data in sensitive or privacy-constrained research environments. Similarly, the strong correlations in *Adult Income* and *Diabetes* datasets validate the utility of synthetic data for predictive modeling and risk factor analysis, where real data may be limited.

A limitation to be addressed is the observable differences at the individual feature level within the SHAP values, which potentially impacts the interpretability of models developed from such data. Moreover, this analysis primarily utilizes mean SHAP values; however, a deeper exploration into the width of SHAP value distributions per variable and the ranking of these variables for feature importance would provide a more nuanced understanding.

The statistical significance of the various correlations presented in the results section, indicated by low p-values, confirms that the observed linear relationships are not by chance, enhancing the credibility of synthetic data in high-stakes research settings. This alignment not only validates the use of synthetic data for training machine learning models where privacy or data availability concerns exist but also underscores the potential of synthetic data to provide meaningful insights comparable to real data.

Chapter 5

Conclusions

5.1 Summary of Contributions

This thesis has made several noteworthy contributions to the field of synthetic data generation and its application in healthcare, underpinned by the development and implementation of innovative optimization strategies that balance data fidelity, utility, privacy, and computational efficiency.

Firstly, we have established notable correlations between fidelity and utility metrics. We demonstrated the usefulness of synthetic data using various healthcare analytics scenarios, focusing on computational efficiency. The HPT strategy employed in this thesis reduces the reliance on utility metrics—indicators of machine learning efficacy—that typically demand substantial computational resources. This efficiency is a pivotal step towards making synthetic data generation more accessible and practicable for a wider range of applications within healthcare analytics and beyond.

Secondly, the adoption of the stratified random sampling method has notably reduced the computational time required for HPT in the Synthetic Data Generation process by up to 75%. This reduction in time and resource expenditure represents a critical advancement in streamlining the process of generating high-quality synthetic data, making it a more feasible option for healthcare organizations of varying sizes and capabilities.

Through various experiments, we have validated the hypothesis that HPT performed on synthetic data is effectively transferrable to real data. This finding opens up new avenues for healthcare organizations to generate synthetic versions of their sensitive patient data and share the synthetic data with external entities for not only building analytics pipelines

but also optimization of predictive models via hyperparameter tuning on synthetic data and using external computing resources. Such a practice not only fosters open data science initiatives within the healthcare sector but also promotes inter-organizational data sharing without compromising patient privacy. Our findings confirm that synthetic data, used for data augmentation to balance class ratios, significantly improves model performance and generalizability. This method effectively addresses data scarcity and imbalance challenges, highlighting synthetic data's role in enhancing model performance and robustness across different contexts.

Our research demonstrates that synthetic data exhibiting higher fidelity metrics—thus ensuring better statistical representativeness—show a stronger correlation between the HPT of predictive models using synthetic and real data. This observation validates our secondary contribution of optimizing synthetic data using Fidelity as the objective function. The implications of this finding are profound, suggesting that the quality of synthetic data, as quantified by its fidelity, is directly linked to the success of hyperparameter tuning and, by extension, the overall effectiveness and reliability of predictive models trained on this synthetic data.

Furthermore, this thesis underscores the potential of synthetic data to catalyze open data science initiatives, extending its impact beyond healthcare to other fields where data privacy and utility are of paramount concern. By enabling secure data sharing between organizations, this work contributes to a broader cultural shift towards collaborative innovation and collective advancement in healthcare analytics.

5.2 Limitations

The evaluation of synthetic data's efficacy in hyperparameter tuning of predictive models has been limited to binary classification and regression (continuous prediction) models. This leaves a significant opportunity to extend the investigation into additional analytical use cases, such as multi-class classification, among others, to provide a more comprehensive assessment of synthetic data's utility in diverse modeling challenges.

This research primarily investigates synthetic data produced via CTGANs and RealTabFormer, pinpointing a focused area within the vast landscape of synthetic data generation (SDG) methodologies. Consequently, this study underscores the necessity for further exploration into alternative SDG techniques to broaden our understanding and application of synthetic data across different contexts.

The study's exploration of use-case agnostic HPT for SDG, focusing on the correlation

between fidelity and utility metrics, requires further experimentation for comprehensive validation. Specifically, optimizing the SDG process using the fidelity metric as the objective function and subsequently examining its correlation across various MLE could enhance the robustness of the findings. This approach will enable a deeper understanding of how adjustments to hyperparameters impact the balance between the fidelity of the synthetic data to the original dataset and its utility across different scenarios.

Regarding dataset scalability, the current study's experiments are confined to datasets containing up to 100,000 records. This limitation suggests that conducting experiments with larger datasets, including real-world hospital data, could offer more robust validation of the hypotheses. Engaging with extensive, real-world datasets from healthcare facilities may unveil nuanced insights into the scalability, efficiency, and applicability of synthetic data methods. This exploration is essential for understanding how synthetic data performs in the context of complex, real-world healthcare data environments, thereby potentially identifying unique benefits or uncovering unforeseen challenges that smaller or less diverse datasets might not reveal.

5.3 Future work

Future research directions will encompass an extensive exploration of privacy metrics, specifically focusing on Membership Inference Attacks and Attribute Inference Attacks. These metrics will play a pivotal role in the multi-objective optimization process during HPT of SDG algorithms. The goal is to achieve a harmonious balance between three critical aspects: fidelity, utility, and privacy of the generated synthetic data. This nuanced approach aims to refine the quality of synthetic data while ensuring stringent privacy protections.

Additionally, the application of HPT strategies to state-of-the-art SDG models, such as diffusion models, will be explored across diverse data types, including time series and multi-relational datasets. This expansion will test the versatility and effectiveness of HPT in enhancing the performance of advanced SDG models across a wider array of data structures, thereby broadening the applicability of synthetic data in complex healthcare scenarios.

Furthermore, targeted efforts will be directed towards refining the capabilities of CTGAN. A significant focus will be on overcoming the challenges associated with generating synthetic data for datasets containing high cardinality categorical variables. Incorporating mechanisms for differential privacy into these models will also be a priority, aiming to bolster the privacy assurances of the synthetic data further. These enhancements are

expected to substantially improve the utility of GAN-generated synthetic data, making it more applicable and reliable for a broader range of healthcare data analytics and research purposes.

Although the data governance for the generation and sharing of synthetic data is out of scope for this thesis, it is of vital importance to establish such policies to ensure the privacy protection and ethical use of synthetic data. Collaborations with healthcare organizations, data custodian agencies, academics and industries are critical to enable the practical validation and refinement of synthetic data generation and sharing within healthcare settings. Engaging with real-world data will not only test the theoretical models in diverse, live environments but also facilitate the translation of research findings into actionable healthcare solutions. This collaborative approach will bridge the gap between synthetic data research and its implementation, potentially leading to significant advancements in the use of AI and ML in healthcare.

References

[1] Theodora Kokosi, Bianca L. De Stavola, Robin Mitra, Lora Frayling, Aiden R. Doherty, Iain Dove, Pam Sonnenberg, and Katie L. Harron. An overview of synthetic administrative data for research. *International Journal of Population Data Science*, 7, 2022.

[2] Nirupa Dattani, Pia Hardelid, Jonathan Davey, and Ruth Gilbert. Accessing electronic administrative health data for research takes time. *Archives of Disease in Childhood*, 98:391 – 392, 2013.

[3] Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (hipaa) privacy rule. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html. [Accessed 01-03-2024].

[4] Danielle Whicher, Mahnoor Ahmed, S Siddiqui, Inez Adams, C Grossman, and Kristin Carman. Health data sharing to support better outcomes. *Washington, DC: National Academy of Medicine*, 2020.

[5] Julia Lane and Claudia Schur. Balancing access to health data and privacy: a review of the issues and approaches for the future. *Health services research*, 45(5p2):1456–1467, 2010.

[6] Adil Hussain Seh, Mohammad Zarour, Mamdouh Alenezi, Amal Krishna Sarkar, Alka Agrawal, Rajeev Kumar, and Raees Ahmad Khan. Healthcare data breaches: insights and implications. In *Healthcare*, volume 8, page 133. MDPI, 2020.

[7] Jan N Van Rijn and Frank Hutter. Hyperparameter importance across datasets. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2367–2376, 2018.

[8] Arunim Garg and Vijay Mago. Role of machine learning in medical research: A survey. *Computer science review*, 40:100370, 2021.

[9] Khushboo Munir, Hassan Elahi, Afsheen Ayub, Fabrizio Frezza, and Antonello Rizzi. Cancer diagnosis using deep learning: a bibliographic review. *Cancers*, 11(9):1235, 2019.

[10] Fabio Alexandre Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. Breast cancer histopathological image classification using convolutional neural networks. In *2016 international joint conference on neural networks (IJCNN)*, pages 2560–2567. IEEE, 2016.

[11] Rahul Paul, Samuel H Hawkins, Lawrence O Hall, Dmitry B Goldgof, and Robert J Gillies. Combining deep neural network and traditional image features to improve survival prediction accuracy for lung cancer patients from diagnostic ct. In *2016 IEEE international conference on systems, man, and cybernetics (SMC)*, pages 002570–002575. IEEE, 2016.

[12] Liya Zhao and Kebin Jia. Deep feature learning with discrimination mechanism for brain tumor segmentation and diagnosis. In *2015 international conference on intelligent information hiding and multimedia signal processing (IIH-MSP)*, pages 306–309. IEEE, 2015.

[13] Ioana Bica, Ahmed M Alaa, Craig Lambert, and Mihaela Van Der Schaar. From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. *Clinical Pharmacology & Therapeutics*, 109(1):87–100, 2021.

[14] Wenbing Chang, Yinglai Liu, Yiyong Xiao, Xinglong Yuan, Xingxing Xu, Siyue Zhang, and Shenghan Zhou. A machine-learning-based prediction method for hypertension outcomes based on medical data. *Diagnostics*, 9(4):178, 2019.

[15] Fatemeh Rahimian, Gholamreza Salimi-Khorshidi, Amir H Payberah, Jenny Tran, Roberto Ayala Solares, Francesca Raimondi, Milad Nazarzadeh, Dexter Canoy, and Kazem Rahimi. Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records. *PLoS medicine*, 15(11):e1002695, 2018.

[16] Kuang-Ming Liao, Chung-Feng Liu, Chia-Jung Chen, and Yu-Ting Shen. Machine learning approaches for predicting acute respiratory failure, ventilator dependence,

and mortality in chronic obstructive pulmonary disease. *Diagnostics*, 11(12):2396, 2021.

[17] Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1):1–10, 2016.

[18] Franck Jaotombo, Vanessa Pauly, Guillaume Fond, Veronica Orleans, Pascal Auquier, Badih Ghattas, and Laurent Boyer. Machine-learning prediction for hospital length of stay using a french medico-administrative database. *Journal of Market Access & Health Policy*, 11(1):2149318, 2023.

[19] Molla S Donaldson, Kathleen N Lohr, and Roger J Bulger. Health data in the information age: Use, disclosure, and privacy—part ii. *JAMA*, 271(18):1392–1392, 1994.

[20] Kah Meng Chong. Privacy-preserving healthcare informatics: a review. In *ITM Web of Conferences*, volume 36, page 04005. EDP Sciences, 2021.

[21] Mark A Rothstein and Meghan K Talbott. Compelled disclosure of health information: Protecting against the greatest potential threat to privacy. *Jama*, 295(24):2882–2885, 2006.

[22] Privacy Act. Personal information protection and electronic documents act. *Department of Justice, Canada. Full text available at http://laws. justice. gc. ca/en/P-8.6/text. html*, pages 4356–4364, 2000.

[23] James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N Cohen, and Adrian Weller. Synthetic data–what, why and how? *arXiv preprint arXiv:2205.03257*, 2022.

[24] Paul Calcraft. What is synthetic data? and how can it accelerate public policy research?, Mar 2022.

[25] Anat Reiner Benaim, Ronit Almog, Yuri Gorelik, Irit Hochberg, Laila Nassar, Tanya Mashiach, Mogher Khamaisi, Yael Lurie, Zaher S Azzam, Johad Khoury, et al. Analyzing medical research results based on synthetic data and their relation to real data results: systematic comparison from five observational studies. *JMIR medical informatics*, 8(2):e16492, 2020.

[26] Ted Laderas, Nicole Vasilevsky, Bjorn Pederson, Melissa Haendel, Shannon McWeeney, and David A Dorr. Teaching data science fundamentals through realistic synthetic clinical cardiovascular data. *BioRxiv*, page 232611, 2017.

[27] Deirdre Hennessy, Claudia Sanmartin, Sabha Eftekhary, Laurie Plager, Jennifer Jones, and Kanecy Onate. Creating a synthetic database for use in microsimulation models to investigate alternative health care financing strategies in canada. *Int J. Microsimul*, 8:41–74, 2015.

[28] Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Stephen Puntis, Rudolf N Cardinal, Angus Roberts, Robert Stewart, and Sumithra Velupillai. Generation and evaluation of artificial mental health records for natural language processing. *NPJ digital medicine*, 3(1):69, 2020.

[29] Mauro Giuffrè and Dennis L Shung. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ Digital Medicine*, 6(1):186, 2023.

[30] Weibin Cheng, Wanmin Lian, and Junzhang Tian. Building the hospital intelligent twins for all-scenario intelligence health care. *Digital Health*, 8:20552076221107894, 2022.

[31] Erica Espinosa and Alvaro Figueira. On the quality of synthetic generated tabular data. *Mathematics*, 11(15):3278, 2023.

[32] Lu Wang, Wei Zhang, and Xiaofeng He. Continuous patient-centric sequence generation via sequentially coupled adversarial learning. In *Database Systems for Advanced Applications: 24th International Conference, DASFAA 2019, Chiang Mai, Thailand, April 22–25, 2019, Proceedings, Part II 24*, pages 36–52. Springer, 2019.

[33] Andrew Jonathan Yale. *Privacy preserving synthetic health data generation and evaluation*. Rensselaer Polytechnic Institute, 2020.

[34] Jerome P Reiter. Using cart to generate partially synthetic public use microdata. *Journal of official statistics*, 21(3):441, 2005.

[35] Haoyue Ping, Julia Stoyanovich, and Bill Howe. Datasynthesizer: Privacy-preserving synthetic datasets. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, pages 1–5, 2017.

[36] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference*, pages 286–305. PMLR, 2017.

[37] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[38] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *arXiv preprint arXiv:1806.03384*, 2018.

[39] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[40] Kieran Chin-Cheong, Thomas Sutter, and Julia E Vogt. Generation of heterogeneous synthetic electronic health records using gans. In *workshop on machine learning for health (ML4H) at the 33rd conference on neural information processing systems (NeurIPS 2019)*. ETH Zurich, Institute for Machine Learning, 2019.

[41] Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P Bennett. Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, 416:244–255, 2020.

[42] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *2016 IEEE international conference on data science and advanced analytics (DSAA)*, pages 399–410. IEEE, 2016.

[43] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019.

[44] Chuan Ma, Jun Li, Ming Ding, Bo Liu, Kang Wei, Jian Weng, and H Vincent Poor. Rdp-gan: Ar\'enyi-differential privacy based generative adversarial network. *arXiv preprint arXiv:2007.02056*, 2020.

[45] Khaled El Emam, Lucy Mosquera, and Richard Hoptroff. *Practical synthetic data generation: balancing privacy and the broad availability of data*. O'Reilly Media, 2020.

[46] Ofer Mendelevitch and Michael D Lesh. Fidelity and privacy of synthetic medical data. *arXiv preprint arXiv:2101.08658*, 2021.

[47] Khaled El Emam. Seven ways to evaluate the utility of synthetic data. *IEEE Security & Privacy*, 18(4):56–59, 2020.

[48] Shanti Gomatam, Alan F Karr, and Ashish P Sanil. Data swapping as a decision problem. *Journal of Official Statistics*, 21(4):635, 2005.

[49] Rudolf Beran. Minimum hellinger distance estimates for parametri models. *The Annals of Statistics*, 5, 05 1977.

[50] Fida K. Dankar, Mahmoud K. Ibrahim, and Leila Ismail. A multi-dimensional evaluation of synthetic data generators. *IEEE Access*, 10:11147–11158, 2022.

[51] Andre Goncalves, Priyadip Ray, Braden Soper, Jennifer Stevens, Linda Coyle, and Ana Paula Sales. Generation and evaluation of synthetic patient data. *BMC medical research methodology*, 20(1):1–40, 2020.

[52] Joshua Snoke, Gillian M Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 181(3):663–688, 2018.

[53] Mi-Ja Woo, Jerome P Reiter, Anna Oganian, and Alan F Karr. Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality*, 1(1), 2009.

[54] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[55] Maher Maalouf. Logistic regression in data analysis: an overview. *International Journal of Data Analysis Techniques and Strategies*, 3(3):281–299, 2011.

[56] Zhongheng Zhang. Introduction to machine learning: k-nearest neighbors. *Annals of translational medicine*, 4(11), 2016.

[57] Yongli Zhang. Support vector machine classification algorithm and its application. In *Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings, Part II 3*, pages 179–186. Springer, 2012.

[58] Yan-Yan Song and LU Ying. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130, 2015.

[59] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[60] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[61] Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, John N Clore, et al. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014, 2014.

[62] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[63] Lei Xu et al. *Synthesizing tabular data using conditional GAN*. PhD thesis, Massachusetts Institute of Technology, 2020.

[64] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. In *Neural Information Processing Systems*, 2019.

[65] Aivin Solatorio and Olivier Dupriez. Realtabformer: Generating realistic relational and tabular data using transformers. *ArXiv*, abs/2302.02041, 2023.

[66] Ann Cavoukian and Khaled El Emam. *De-identification protocols: essential for protecting privacy*. Information and Privacy Commissioner of Ontario, Canada, 2014.

[67] Gregory E Simon, Susan M Shortreed, R Yates Coley, Robert B Penfold, Rebecca C Rossom, Beth E Waitzfelder, Katherine Sanchez, and Frances L Lynch. Assessing and minimizing re-identification risk in research data derived from health care records. *eGEMs*, 7(1), 2019.

[68] Treasury Board of Canada Secretariat. Directive on privacy impact assessment, Aug 2017.

[69] Health Canada. Government of canada, Apr 2019.

[70] Philipp Probst, Anne-Laure Boulesteix, and Bernd Bischl. Tunability: Importance of hyperparameters of machine learning algorithms. *The Journal of Machine Learning Research*, 20(1):1934–1965, 2019.

[71] Bernd Bischl, Olaf Mersmann, Heike Trautmann, and Claus Weihs. Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evolutionary computation*, 20(2):249–275, 2012.

[72] Li Yang and Abdallah Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316, 2020.

[73] Radwa Elshawi, Mohamed Maher, and Sherif Sakr. Automated machine learning: State-of-the-art and open challenges. *arXiv preprint arXiv:1906.02287*, 2019.

[74] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. *ArXiv*, abs/2209.15421, 2022.

[75] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE transactions on neural networks and learning systems*, PP, 2021.

[76] Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. *ArXiv*, abs/2210.06280, 2022.

[77] Khaled El Emam, Lucy Mosquera, and Jason Bass. Evaluating identity disclosure risk in fully synthetic health data: Model development and validation. *J Med Internet Res*, 22(11):e23139, Nov 2020.

[78] Joao Fonseca and Fernando Bacao. Tabular and latent space synthetic data generation: a literature review. *Journal of Big Data*, 10(1):115, 2023.

[79] Daniel Mesafint Belete and Manjaiah D Huchaiah. Grid search in hyperparameter optimization of machine learning models for prediction of hiv/aids test results. *International Journal of Computers and Applications*, 44(9):875–886, 2022.

[80] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

[81] Shuhei Watanabe. Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. *arXiv preprint arXiv:2304.11127*, 2023.

[82] Fida K Dankar and Mahmoud Ibrahim. Fake it till you make it: Guidelines for effective synthetic data generation. *Applied Sciences*, 11(5):2158, 2021.

[83] Zdravko Botev and Ad Ridder. Variance reduction. *Wiley statsRef: Statistics reference online*, pages 1–6, 2017.

[84] Nadia Burkart and Marco F Huber. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317, 2021.

[85] Alexandr Oblizanov, Natalya Shevskaya, Anatoliy Kazak, Marina Rudenko, and Anna Dorofeeva. Evaluation metrics research for explainable artificial intelligence global methods using synthetic data. *Applied System Innovation*, 6(1):26, 2023.